Wittgenstein and the Concept of Learning in Artificial Intelligence

Wittgenstein og begrepet om læring i kunstig intelligens

Arturo Vázquez Hernández

Main Supervisor: Alois Pichler
Co-supervisor: Simo Säätelä

A thesis presented in partial fulfilment of the requirements for the degree of Master of Philosophy

Department of Philosophy
University of Bergen (UiB)
Norway
Spring Semester 2020
FILO350

A B S T R A C T

The object of this investigation is to analyze the application of the concept of learning to machines and software as displayed in Artificial Intelligence (AI). This field has been approached from different philosophical perspectives. AI, however, has not yet received enough attention from a Wittgensteinian angle, a gap this thesis aims to help bridge. First we describe the use of the concept of learning in natural language by means of a familiar and of a less familiar case of human learning. This is done to give us a general idea about the meaning of this concept. By building two basic machine learning algorithms, we introduce one of the technical meanings of "learning" in computer science, i.e. the use of this concept in machine learning. Based on a study and comparison between both uses, the one in ordinary language and the one in machine learning, we conclude that both usages exemplify one and the same "family resemblance" concept of learning. We apply this insight further in a critical discussion of two specific philosophical positions about the applicability of psychological or mental concepts to software and hardware, especially in AI. One of the contributions of this investigation is that the use of mental concepts concerning machines does not imply the ascription of a mind.

S A M M E N D R A G

Målet for denne undersøkelsen er å analysere anvendelsen av begrepet om læring på maskiner og programvare slik de blir fremstilt innenfor Kunstig intelligens (KI). Flere forskjellige filosofiske perspektiver har tatt for seg dette feltet. KI har derimot ikke fått tilstrekkelig med oppmerksomhet fra et wittgensteinsk perspektiv. Dette er et hull denne masteroppgaven prøver å fylle. Først beskrives bruken av begrepet om læring i naturlig språk både i en fortrolig og en mindre fortrolig situasjon av menneskelig læring. Dette er gjort for å gi leseren en generell forståelse av meningen med dette begrepet. Ved å lage to grunnleggende maskinlærings-algoritmer blir en av de tekniske betydningene av «læring» i informatikk introdusert, dvs. bruken av dette begrepet i maskinlæring. På bakgrunn av et studie og en sammenligning mellom de to bruksmåtene – den hverdagslige og den sistnevnte fra maskinlæring - konkluderes det at begge bruksmåter er del av det samme «familielikhets»-begrepet om læring. Denne innsikten anvendes videre i en kritisk diskusjon av to spesifikke filosofiske posisjoner om anvendelsen av psykologiske begrep i programvare og maskinvare, spesielt i KI. En konklusjon blir da at anvendelsen av psykologiske begrep på maskiner ikke impliserer at de tilskrives et sinn.

Dedication

To my parents and my brother,
for their unconditional support, patience, and love.
Without them, none of this would have been possible.

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Acknowledgements

Writing this thesis has been a personal and academic challenge. I want to express my gratitude to the people that helped me throughout this process. To Alois Pichler, for taking the time to be my supervisor. For his friendship, guidance, and encouragement in these two years. To Simo Säätelä, my co-supervisor, for his welcoming disposition to read and comment my writings, and for giving me the opportunity to collaborate in the Nordic Wittgenstein Review. To them I owe being introduced to the field of philosophy that I find the most fascinating. To Pekka Parviainen, for his kind advise in machine learning, and for his comments and corrections in the third chapter. To my professors Ole Hjortland and Sorin Bangu for their suggestions at the end of my investigation. To Deirdre Smith, Steinar Thunestvedt, and Kirsten Bang, for their institutional support and their commitment to the student community. Lastly I want to thank the Department of Philosophy, for providing the tools, resources, and workspace I needed to conclude this project. Thank you.

Contents

List of Figures

Abbreviations of Wittgenstein's Works

BB	Blue and Brown Books (Wittgenstein 1969)
BT	The Big Typescript: TS 213 (Wittgenstein 2005)
CE	Cause and Effect: Intuitive Awareness (Wittgenstein 1993)
LFM	Wittgenstein's Lectures on the Foundations of Mathematics, Cambridge 1939 (Wittgenstein 1976)
OC	On Certainty (Wittgenstein 1991)
PI	Philosophical Investigations (Wittgenstein 2009a)
PPF	Philosophy of Psychology: A Fragment (Wittgenstein 2009b)
RFM	Remarks on the Foundations of Mathematics (Wittgenstein 1983)
RPP I	Remarks on the Philosophy of Psychology I (Wittgenstein 1980a)
RPP II	Remarks on the Philosophy of Psychology II (Wittgenstein 1980b)
Z	Zettel (Wittgenstein 1967)

References to **BB**, **CE**, **LFM**, are to the pages of the cited editions. References to **BT**, **OC**, **PI**, **PPF**, **RFM**, **RPP I**, **RPP II** and **Z**, are to the sections and paragraphs.

Chapter 1

Introduction

1. Ludwig Wittgenstein's so-called later thought has a distinctive ambivalence: it is considered to be, alongside Plato and Kant, one of the main contributions to philosophy in general and, at the same time, it is one of the most avoided by contemporary philosophers. Almost everyone acknowledges that Wittgenstein's contributions to philosophy are of the utmost importance. Even though numerous contemporary philosophers do recognize the significance of this perspective, they do not confront their ideas with the criticisms posed by this way of thinking.

One of the various reasons for this overlooking is that some philosophers think that Wittgenstein's thought represents in some sense the end of philosophy. The family of methodologies proposed by this author, they think, leads to the conclusion that philosophical problems are not real problems, but merely linguistic confusions. The depth and sublimity of philosophy, its foundational character, have been replaced by a series of considerations about how to use certain terms to prevent conceptual misunderstandings. If this is the case, some people might think, the only thing we have left is to either repeat/reinterpret what Wittgenstein said or to simply ignore it.[1]

2. Contrary to this spirit, what I think is one of the main contributions of Wittgenstein's philosophy is that we can apply it: we can make use of the different methodologies of this philosophy to extract significant clarifications from particular practices. Hacker has maintained, for example, that the *Investigations* represents a Janus-faced book: its first part can be considered to be negative in the sense that it poses a critical analysis of the *Tractatus* and of a robust portion of analytic philosophy. The second part, in contrast, shows positive clarifications concerning several philosophical issues from a new angle (Hacker 2010, p. 277). By using a series of methods in the *Investigations*, Wittgenstein applies this perspective mainly in two fields: mathematics and psychology. Let me offer two examples.

Some philosophers take expressions like "9", "+" and "π" as the names of certain entities or objects. The natural questions here are "What do they mean?", "These concepts are in which object's place?" Mathematical Platonism, for example, holds that these concepts are names of abstract objects that exist independently of us and our language. Similarly, in philosophy of mind concepts like "pain", "sensation" and "thinking" sometimes are taken as names standing in the place of some object.

[1]For a deeper treatment of why Wittgenstein's later thought has been ignored since the second part of the twentieth century by philosophers of the analytic tradition, particularly in Britain and the United States, see Tripodi 2020.

Philosophers that support this perspective argue that these concepts point to, or are names of, private experiences. Moreover, these experiences can be directly apprehended only by the subject of experience through introspection, a process which allows us to name them.

From a Wittgensteinian perspective, in contrast, we do not consider that the meaning of these concepts are objects, neither abstract in the case of mathematics nor private in the case of psychology. The meaning of these concepts is not a something for which they stand. This is not to say, of course, that they do not have meaning. Their meaning, just like that of almost every other word, is their use in natural language (PI, 43). This implies, among many other things, that we cannot define these concepts in terms of necessary and sufficient conditions of application: their uses are not decided in advance. They are diverse and depend on each context. These concepts have different meanings in different circumstances, without losing certain regularity and similarity in their applications.

With these methodological remarks in mind, defining "9" as the name of an abstract entity would impede us to see its different applications, that in a definite context is used to count (PI, 1). This may sound trivial but consider the importance of the practice of counting in our life. The meaning of "9", of course, is not exhausted by this usage. Even though the roles of our words are not so evident they can be as diverse as the functions of the objects in a toolbox (PI, 11). Likewise, the analysis of our psychological concepts shows that their use presents an important asymmetry between the first and the third person. Expressions like "Now I understand the procedure!" are declarations and not reports of private mental processes. Conversely, expressions like "He has understood how to proceed!" are descriptions that sometimes are related to complex criteria (Z, 472).

A similar contribution is offered in *Philosophical Foundations of Neuroscience* (Bennett and Hacker 2003). This work represents another example of what I like to call, *applied Wittgensteinian philosophy*. By distinguishing the domain of the empirical from the domain of the conceptual, both authors detected important misunderstandings in cognitive neuroscience and clarified many of the concepts that are used in this practice, for example, introspection, sensation, perception, memory, consciousness, and self-consciousness. They explored in-depth some issues that lie in the conceptual core of this science: the mereological fallacy, the problem of other minds, the mind-body problem, qualia and conscious experiences, etc. Although these philosophical views are not exempt from examination and controversy, they represent significant threats for the theories they criticize, by exhibiting important misunderstandings and vulnerabilities in the heart of their conceptual frameworks.

Based on the examples given above, Wittgenstein's thought is not a dead-end track. It does not represent the finale of philosophy. Rather, this way of thinking offers tools and methodologies for clarifying many of our contemporary practices: philosophical, scientific, artistic, religious, political, etc. Wittgensteinian philosophy, apart from being an important contribution to philosophy itself, can be applied, arguably, outside the purely philosophical field. Because language is the medium of science, political organization, and culture, no human practice is immune to conceptual problems and misunderstandings. The malleability of Wittgenstein's philosophy enables us to think critically —from the perspective of language and the use of concepts— about forms of representation and contemporary issues that shape our world.

3. In this sense it is possible to give a positive account of Wittgensteinian philosophy. The main interest of this work is to apply this family of methodologies to Artificial Intelligence, so we can achieve a better understanding of it. AI cannot be defined easily. There is no single feature that identifies every aspect of this group of practices. Broadly speaking, AI can be understood as the development of machines capable of carrying out tasks which, in the case of humans, require intelligence. However, in closer examination this is much more complex. In AI a multitude of disciplines converge, such as symbolic logic, computer engineering, applied mathematics, and neurophysiology, for example. AI consists of several techniques, methods, software, machines, and applications that address different problems and tasks, by processing different kinds of information. From the internet to stock markets, from science to language translation apps, AI is used in most parts of our contemporary lives.

This investigation explores the use of some of the elementary concepts in AI. In general, we are concerned with *the application of psychological* or *mental concepts to software*. To develop technology that works intelligently, AI has been divided into many areas, among them, for example, natural language processing, knowledge representation, perception, artificial general intelligence, and machine learning.[2] In particular, we will focus on *the use of the concept of learning* in the last of these fields. To offer a comprehensive account of this concept in machine learning, first we need to bear in mind that *learning* is primarily a concept from our natural language. Thus, a comparison between the application of the concept of learning to software and its application in natural language, i.e. concerning human beings, is needed. By shedding light on both uses of the concept of learning, this analysis will enable us to advance philosophical considerations about the sense and features of the application of mental concepts to software. Needless to say, our aim is not to criticize the progress machine learning has displayed in the past decades. The technological and scientific advances machine learning has fostered are beyond doubt. Our investigation is about the legitimacy of the use of some of the central concepts in this practice.

4. The structure of this work is as follows. To properly understand some of the concepts of our natural language it is necessary to describe their use in familiar and less familiar cases. With this in mind, in the second chapter we offer a description of the meaning of learning as a family resemblance concept —a notion which is crucial in advancing this investigation. First we describe a well-known, familiar case of learning. This type of case refers to situations in which we would say, for instance, that someone is learning to play chess, learning to solve mathematical equations, or learning to play the bass under the supervision of an expert or an intellectual authority. In addition, we describe one of the less familiar examples of learning, i.e. a case of autodidacticism or self-teaching, like Pascal, Kubrick, and Borges. Arnold Schönberg, for instance, started to compose music for the violin at the age of nine without supervision. We say that these people can teach themselves how to perform a certain task without instruction.

The concept of learning has received direct attention from empirical fields such as psychology, pedagogy, and neuroscience. In Wittgenstein scholarship, however,

[2]For practical purposes, in this work we understand machine learning as a subfield of AI. Nevertheless, some computer scientists think that machine learning deserves to be considered a separate field of research.

this concept has not had the same fortune. Regarding the domain of the mental, philosophers mainly deal with so-called higher-order concepts, like understanding, thinking, or meaning. This chapter is relevant because it offers the grammar of learning, not as a subordinate concept, but as a concept worth of having a philosophical consideration of its own.

The third chapter deals with the use of *learning* in machine learning. Due to the large extent of approaches, methods, applications, and problems that constitute this practice, this investigation is focused only on two machine learning algorithms, i.e. supervised and unsupervised learning. Such approaches have been chosen because of the relation they seem to have with the instances of human learning that are analyzed in the second chapter. We will introduce one imaginary case for supervised and another one for unsupervised learning. In these sections we describe how these machines work —i.e. how they receive and process data, the mathematical methods they use to analyze the data, and the outcomes they offer— in order to show how experts apply mental concepts to them. This description will make it easier for us to understand expressions like "The machine learned the patterns in the data", and others.

After human learning and machine learning have been described, in the fourth chapter we provide a comparative analysis between them, emphasizing some similarities and differences. At the end of the chapter, a comparative table of the different uses of learning is provided. By seeing how these uses work, in the fifth chapter we ask for the sense of certain applications of mental concepts to software under particular circumstances. This, in turn, will allow us to offer solutions to some philosophical problems regarding mental concepts in connection with machines, like strong artificial intelligence —the view according to which machines can be bearers of mental concepts in the same way humans are— and perspectives that discard, or consider as meaningless, ascriptions of mental concepts to software. One contribution of this investigation is the claim that the ascription of mental concepts to machines does not imply the ascription of a mind. Nonetheless, the family resemblance structure and grammar of mental concepts allow us to apply them beyond human beings and intelligent animals, and thus, the application of psychological expressions to software can be legitimate and meaningful under certain circumstances.

Chapter 2

The Grammar of *Learning*

The general objective of this chapter is to describe different uses of the concept of learning in natural language. For this, we show that *learning* is a family resemblance concept and has manifold applications and uses. Among the different uses of this concept we distinguish between familiar and less familiar applications. To achieve a better understanding of this concept, it is proposed that both types of uses must be described. On this basis, we introduce two language-games, the first of which exemplifies the familiar use of *learning* and the other illustrates its less familiar use. Similarities and differences between both applications are shown.

The relevance of this chapter resides in the fact that the concept of learning has been considered in depth mainly by psychology, pedagogy, biology, and neuroscience, among other fields. In Wittgenstein scholarship, however, it seems that this concept has not had the same scrutiny as some of the so-called "higher-order concepts", like understanding, representation, thought, and intentionality (see Williams 1999a and Weber 2019). As this chapter aims to show, the concept of learning plays a crucial role in Wittgenstein's philosophy and methodology, and deserves further philosophical consideration.

2.1 *Learning* as a family resemblance concept

Among the vastness and complexity of our language, we can trace a distinction between two groups of concepts. The first refers to concepts that can be learned by means of a general definition, while the second one refers to concepts that cannot be learned in this way.[1] Definitions are one of the landmarks of science and, as such, they contribute to increase our knowledge. Let me offer an example.

In geometry, a hypotenuse is defined as the longest side of a right triangle. This definition states the necessary and sufficient conditions for the concept. Being a side is a necessary condition for something to be a hypotenuse. This implies that we can ascribe this concept only to sides —and not to angles. However, being a side is not a sufficient condition for something to be a hypotenuse, because the side might be a cathetus. The sufficient condition is being the largest side of a right triangle. This definition implies that only if these features take place, then having

[1]This distinction is of course neither exhaustive nor exclusive. As Wittgenstein points out: "how we group words into kinds will depend on the aim of the classification – and on our own inclination" (PI, 17).

a hypotenuse is a guarantee. If something does not have these features, then this is not a hypotenuse.

Similarly, we can find several definitions in mathematics and science. Think for example about the definition of "space" in the theory of relativity or Euclidean geometry, the definition of "atom" or "water" in chemistry, of "organism" in biology, etc. In general, a definition points to the essence of the concept in question, allowing us to distinguish such concept from others. The definition —being the outcome of a scientific discovery or being the cornerstone of a particular system—, we might say, prepares the terrain for acquiring knowledge by means of advancing a scientific investigation.[2] By having established the definition of hypotenuse, for example, we can make calculations and predict certain states of affairs in the most diverse fields, such as astronomy, optics, navigation, etc.

However, not all of our concepts can be learned by means of a definition. The second group contains what Wittgenstein called *family resemblance concepts*. Wittgenstein introduces the notion of family resemblance in opposition to essentialists, who think that there is an essential feature, a common characteristic that gives an account of why the different instances of a concept "fall under it" (Glock 1996, p. 120). In this sense, this term presents an alternative to definitions and aims to show how a large number of concepts work. Wittgenstein exemplifies this notion with the concept of game. Imagine that someone proposes the following definition: "Playing a game consists in moving objects about on a surface according to certain rules" (PI, 3). We could say that this person is only considering activities like checkers or chess. Her definition, thus, ignores pétanque, football, and Sudoku, among many other games. Someone might think that, given this shortcoming, our task is to rectify or expand such definition to include the omitted games. But if we look at the diversity of games we have, it is not possible to offer a definition that includes all of them. There are countless kinds of games. No matter how general the definition is, it would explain only a portion of the word's applications, while excluding others. The uses of some concepts, such as *game*, are dynamic and change over time. They are not decided in advance.

Wittgenstein invites us to look at the different games we have. The word "game" has a variety of uses in different contexts. We speak of amusing games, games that are difficult to understand, others where luck is a central element, games that require physical skills like tennis, video games, games with numbers, games that involve precision such as carom billiards, others that require long calculations like Go, games that demand good mental skills like memory, ball games, card games, horse racing, auto racing, board games, games that children invent in leisure time, games played in ice, etc. If we compare all these games, we will find some affinities, for instance, that some are based on competition, or that some are entertaining. But these similarities vanish when we consider other games. There is nothing essential and necessary that runs through all these games and allows us to use one single concept to contain them all. Instead, our games form a family, a complex "network of similarities overlapping" (PI, 66), in which some of its members share features and resemble each other. *Game*, thus, is a family resemblance concept.

In the *Investigations* Wittgenstein offers several examples of family resemblance

[2]This does not mean that definitions belong exclusively to science. In our daily activities we use definitions regularly. Consider the definition of "guitar", of "checkmate" in chess, of "bourgeoisie" in Marxist political philosophy, etc.

concepts, such as language, proposition, description, name, rule, number (PI, 1-66), and some mental or psychological concepts like remembering, wishing, recognizing (PI, 35), imagining, understanding (PI, 135ff, 363ff), etc. A family resemblance concept does not have an essential core. There is no one thing in common among its different uses which legitimizes its application in different circumstances.[3]

This, I believe, also applies to the concept of learning. To illustrate a small portion of the countless instances of human learning, let us offer some examples extracted from the *Investigations*. We speak of learning:

- a native language (PI, 1)
- a second language (PI, 32)
- the meaning of a word (PI, 1)
- to bring an object at a particular call (PI, 2)
- to talk (PI, 5)
- the numbers by heart (PI, 9)
- the colors (PI, 28)
- to play a game (PI, 31)
- to apply a certain rule (PI, 54)
- the use of concepts in ethics and aesthetics like *good* or *beautiful* (PI, 77)
- to read tables and charts (PI, 86)
- to read in the native language (PI, 156)
- foreign expressions (PI, 159)
- algebra (PI, 179)
- the meaning of names of sensations, like *pain* (PI, 244)
- to lie (PI, 249)
- the use of the word *thinking* (PI, 328)
- to calculate (PI, 385)
- from experience (PI, 315, 354)

Some of these uses have similar features, but these very features disappear when we consider other instances of learning. Think about the similarities and differences among the uses above. Some of the instances of human learning are related to different kinds of training and instruction. Compare the circumstances in which we learn to calculate in school and the circumstances in which we learn to play a game, say, in leisure time. Different kinds of capacities are required. Consider the role of memory in learning a second language and the role of pretending in learning to lie. Saying that someone is learning to talk is different from saying that someone is learning to read. The criteria for using the word in each case are different. We test the learner's progress in different ways, for instance, by means of exams in the case of learning algebra, and by actually bringing up the learner to play in the case of chess. Think about the goals of each practice. What is the purpose of learning to bring an object at a particular call? It may be the construction of a wall. Compare this aim with the possible purposes of learning the use of a new word —it is also worth noting that when learning a new word, we do not normally learn to describe its use (Z, 114). Additionally, we might say that in some instances of learning the subject is aware of the process of learning something. This last point does not apply

[3]Although games, for example, are activities, we do not call them "games" *because* they are all activities. Being an activity is not enough for something to be a game. See Forster 2010, p. 69 and Glock 1996, p. 121.

to the case of learning a native language or learning the use of some basic concepts like *pain*.

The circumstances in which we apply the concept of learning are manifold. We use this concept in a variety of ways, and there is no single feature common to all of its applications in natural language. Some of them resemble each other by sharing features, such as 'learning by repetition and exercise', similar kinds of training and instruction, the role of certain mental skills such as memory, etc. What if someone replies: "But there is a common feature to all of these instances. One always learns *something*." In natural language, however, "I am learning something" is a proposition that demands further questions, namely, "What are you learning?" "What do you mean by *something*?" Paradoxically, 'learning something' is not learning, and it is not a proposition of our natural language.[4]

This of course does not mean that no definition of *learning* is possible or useful. Within the different applications of a family resemblance concept, it is possible to draw rigid boundaries for special purposes (PI, 68). This, for example, is the case of —at least some— definitions in science. Consider the concept of number. Although it is not possible to define this concept by giving necessary and sufficient conditions, we still can establish borders to define a particular region of applications for the concept in question. We have natural and real numbers, rational and irrational numbers, negative numbers, prime numbers, complex numbers, etc. Some of the different uses of the concept of number resemble each other. But there is no single feature that all of them share which legitimizes our use of "number". However, the natural numbers have been defined, for instance, as the numbers we use to count, or more technically, as the integers greater than zero.

In this sense, empirical research has been deeply concerned with the concept of learning. Neuroscience, biology, and psychology, for example, have contributed to our knowledge of how human beings and other species learn. Some contemporary behaviorist approaches, for instance, have defined "learning" as being a "biological mechanism. [...] It is first and foremost a survival mechanism, a means of meeting the challenges that threaten our survival." Therefore, learning "did not evolve so that [we] could learn to solve algebra, word problems or program a computer" (Chance 2014, p. 1). This means that *learning* is understood essentially as a set of abilities we are naturally equipped with that makes it possible for us to adapt to new environments (see Chance 2014, p. 25). This field is concerned with explaining the causes of adaptive behavior based on natural selection, as well as detecting similarities between the behavior of humans and animals. Based on empirical research, then, this definition provides useful grounds to answer scientific questions. By focusing on this definition, psychologists need not be concerned with the more complex concept of learning of natural language (see Kuusela 2013, p. 63).

However, the former is a derivative concept of learning. Its use is scientific and aims to address only empirical problems. As such, this characterization cannot offer an account of all the instances of learning. As a family resemblance concept, learning

[4]The expression "One always learns *something*", at the most, can be regarded as a grammatical proposition of the concept of learning. This means that this proposition shows how the concept of learning is used. It does not provide us with information about the concept. Compare these remarks with the expression "One always knows *something*." Both expressions show that these concepts are used transitively. We must note that in the case of *learning* we also say things like "She learns quickly." By using this proposition we express something about the behavior, actions and skills of someone.

has an open-ended number of applications. Thus, in order to have a comprehensive grasp of this concept's multifariousness, we need to describe its uses, some of which can be more or less familiar than others.[5]

It is worth noting that the distinction between familiar and less familiar instances of learning is not something fixed. The aim of this distinction is not normative. We are not trying to delineate the border between different applications of our concept. Language changes. Our concepts and forms of representation are not static. Nevertheless, it is possible to identify applications that could be labeled or described as familiar or usual within the different uses of this concept, for example, "The child is learning to count". This can be regarded as a typical, familiar example of learning, not because it represents *the essence* of learning, but merely because this is a well-known situation. Similarly, we can recognize expressions that we would not qualify as typical examples of the uses of this concept, such as "She learned the value of responsibility". This last case is indeed an instance of learning, but it does not share some of the features with typical or more familiar applications of the concept. Being responsible does not mean doing this or that. This concept is related to how we do something and the reasons for doing it. Thus, learning this value involves a different —and perhaps a more complex— scenario, where other skills are required, other criteria are applied, etc. The description of familiar and less familiar cases, then, is necessary to have a better understanding of the concept (PI, 130). If we proceed to analyze only the more familiar cases, some features of the concept of learning may be overlooked.

In the following we introduce two language-games in which the concept of learning is used. The first one represents a familiar situation that will provide insight into the more typical instances of human learning. The second language-game, the less familiar use of *learning*, is not to be considered an exception or counterexample to the first one, but as another legitimate, though uncommon, case of learning, one that complements the family of uses of the concept. This language-game will help clarify that the concept of learning has countless applications related by family resemblances.

[5]I take this distinction from a conversation I had with Martin Gustafsson on the concept of understanding. Some may think that understanding an order means to obey it. This, we could say, represents a familiar instance of understanding: if someone proceeds according to the rule, this means that she understands the rule. However, if this is taken as the definition of understanding, we would endorse a behaviorist-type approach that does not give an account of many instances of understanding. There are cases in which, say, a soldier understands the order and does not obey it. Perhaps the soldier panicked. These cases can be seen as less familiar, because the understanding is manifested in the possible responses (reasons) the soldier can give about why she didn't comply. To give a comprehensive account of a family resemblance concept, then, we must describe both familiar and less familiar instances of the concept in question.

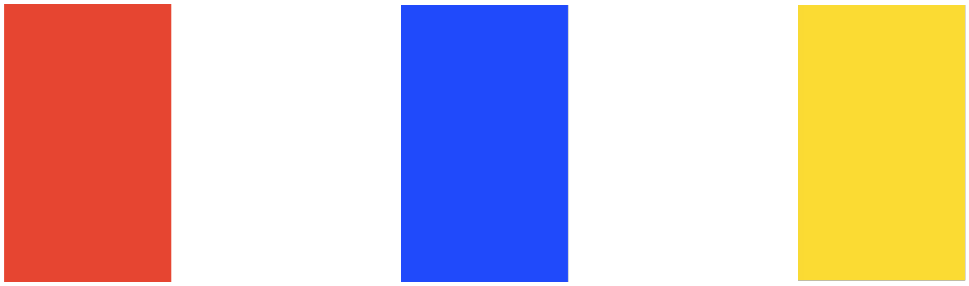
2.2 Two language-games of *learning*

Let us ask first, regarding the concept of learning, which are the expressions we find the most common and familiar? Under which circumstances do we ordinarily apply the concept of learning? This concept has manifold uses in our everyday life, and one of its most natural applications, I think, concerns infants. We gradually introduce them to our practices: they learn to talk, to count, to walk, to play, etc. Let us explore one case in which the use of the concept of learning is clear and non-problematic (see PI, 5).

2.2.1 The language-game of color learning

How does a child learn the colors? An answer to this question, once again, does not require a scientific, neurological account of this process. Rather, we aim to describe the circumstances in which we would use expressions like "The child is learning the colors" and similar ones.

Let us construct an imaginary scenario. In a kindergarten, a teacher prepares some tools to introduce children to the primary colors. She places three sheets of paper in front of the class, a red, a blue and a yellow one, so everyone can see them. See Figure 2.1.



(a) Red (b) Blue (c) Yellow

Figure 2.1: Didactic material 1

"We will learn the colors today", she tells them. While pointing to each of the sheets of paper in turn, the teacher names the corresponding colors. She utters: "*This* color is red, *this* one is blue, and *this* color is yellow." The teacher invites the children to repeat the name of each color after her. After some training, children start to be familiar with the color-words. To test the children's understanding, the teacher gives the following order: "Find something red in the classroom, like *this* apple over here!", while showing the fruit to the children as an example of a right course of action —we can imagine that for this activity the teacher placed red things all over the classroom the day before. An important part of the training lies in continuing uttering the color-word, in this case "red", in connection with the object-word. The teacher wants the children to associate the name of the color with the object in question. We can imagine one of the pupils saying, "I found a red t-shirt!" while pointing to the garment she herself is wearing. If her use of words is correct, that is, if the t-shirt she is wearing is indeed red, the teacher will encourage this behavior by showing approval and satisfaction. If the pupil's use of "red" is incorrect, it is expected that the teacher reacts differently, perhaps expressing disagreement "No! *this* is not red. Red is the color of *this* ball!" while

holding the object in her hands, showing another example. After some training, the teacher may hang in the wall pictures of red objects containing the tag "red" in them, so the students can associate the object, the color, and the word "red" — This may contribute to the children's learning to spell the color-word. See Figure 2.2.

Figure 2.2: Didactic material 2

After the teacher thinks that the children are successfully acquiring the concept of red, she may repeat this process with the other two colors. If someone asks her "What are your students doing?" she may respond "They are learning the colors." In this and similar ways we use this expression.

Let us analyze this use of words. "By saying that the children are learning the colors, does the teacher mean that an internal process is happening inside the students' minds? Does *learning* mean a cognitive process inside the children's brains?" —Not in this situation. The teacher uses this expression having in mind the training she gives to the students, the children's behavior toward the didactic material she provides in class, and the abilities they are gradually acquiring.[6] "Can we determine the exact point in time when the child in fact learns to apply the color-word 'red'?" —This is not how the concept of learning is used. It is not possible to establish sharp boundaries in the process of acquiring an ability —and this means *we do not do it*. "11:43 in the morning is the exact time in which the student learned to use the word" is not a meaningful expression. This expression has no use in our natural language. Instead, the meaning of these words is established gradually, to the point where children "can reliably use [such words] in new contexts" (Weber 2019, p. 693).

The expression "The student learned (is learning) the colors" is used in certain circumstances, for example, while testing the children's understanding with questions like "What color is *this*?" while pointing to a certain object. As Nelson argues, "we attribute concepts to individuals when a 'non-deviant' use of an appropriate word is observed in the appropriate context" (Nelson 2009, p. 278). This means that a student learns to use a certain color-word when there is uniformity in her applications within the practices of the community. She succeeds when her uses agree with ours. It is worth noting that repetition is of fundamental importance in this process. At the beginning of the process of learning a new word, we can imagine the student repeating without understanding what the teacher utters. As long as the training progresses, children gradually acquire the concepts from applying

[6]Even if a certain process goes on in the student's brain when learning x, such process is not sufficient for us to say she learned x. We say it until the student displays it in action. Correct behavior is both necessary and sufficient for us to use this concept. "'He understands' must have more to it than: the formula occurs to him. And equally, more than any of those more or less characteristic *concomitant processes* or manifestations of understanding" (PI, 152). Empirical fields, such as neuroscience, ask for the concomitant processes and "the neural processes that make it possible for animals to [learn] and remember whatever they can [learn] and remember" (Bennett and Hacker 2003, p. 154).

(correctly) certain words. Put differently, we could say that in this scenario children acquire first the color-words and then the concept of red, and not backward.[7]

Based on its use, the concept of learning presents a variation. The meaning of this concept is slightly different depending on how and in which tense we employ it. Let us take Wittgenstein's remarks on the concept of understanding as a benchmark. In the discussion of rule-following Wittgenstein shows that the grammar of the concept of understanding is "closely related" to the grammar of "being able to" (PI, 150). Imagine going to an event with someone who knows French, and by coincidence there is a French lady that does not speak any other language. At that moment we may point to our friend while saying something like "She understands French! She can translate what the French lady is saying." Although this concept is much more complex, in this case, understanding is akin to have an ability, to be able to do something.[8]

The use of learning in the past tense is similar to the concepts above. Expressions like "The student learned the colors" suggest that she has mastered the technique, that she already understands the system. In contrast, expressions in the present tense like "The child is learning the colors" indicate that the student is in the process of mastering or acquiring an ability, or coming to understand a system (PI, 143). In the process of ability acquisition the student is expected to err (at least more frequently). Consider the following remark.

> "Learning [something]" presumably means: being brought to the point of being able to do it. (PI, 385)

Used in the present tense, thus, the concept of learning is understood as a process extended in time, in which the learner becomes gradually "proficient in the use of a particular word" (Nelson, p. 280, 2009). This means that the number of mistakes decreases as the process goes on. The conceptual network of "learning" and "process" consists of concepts like time, practice, training, repetition, error, among others.

In connection with the above, consider the fact that at least in this language-game, the grammar of learning suggests that children do not learn by explanation. The teacher is not explaining what a color is. She is not lecturing children about the nature of colors. The method, or the series of methods for the students to learn are mainly didactic and involve actions and utterances. The gist of this language-game is not to teach the theory of color, but to teach the student to use the color-words in question. This allows the student to acquire certain abilities and allows us to introduce her to our practices, in particular to the network of language-games involving colors and related ones, until she can be regarded as a competent user of this complex vocabulary like any one of us. In many passages regarding concept

[7]The discussion about concept formation in Wittgenstein's later philosophy has attracted the attention of many philosophers and scientists. In my opinion, the expression *concept formation* is unfortunate, for it intimates an internal process. "Concept attribution", I think, stresses the fact that a correct application of an expression depends on public criteria and not on the individual's brain/mental processes. See RFM VI, 7-8: "He tells us: 'I saw that it must be like that.' [...] This *must* shews that he has adopted a concept."

[8]This point is intimately related to the knowing-that and knowing-how distinction in *The Concept of Mind* (Ryle 2009). Ryle has a similar approach to the concept of understanding. He also stresses the importance of action in the process of learning and acquiring different abilities. For further development of the different uses of *understanding*, see Pichler 2018.

formation and language learning Wittgenstein is concerned to attract our attention towards this point. Consider the following remarks.

> Here the teaching of language is not explaining, but training. (PI, 5)
>
> Any explanation has its foundation in training. (Educators ought to remember this.) (RPP II, 327)

Another important point of our language-game is that the authority figure has a central role. Without a guide, children would not be able to learn the colors, to read, to talk, etc. As Hardwick notes, the child learns "to associate a word with an object or activity by means of encouragement given him by his teacher or parent" (Hardwick 1971, p. 95). When teaching new concepts, Wittgenstein notes that we make use of examples and exercises to guide children's behavior and to foster certain reactions: "I do it, he does it after me; and I influence him by expressions of agreement, rejection, expectation, encouragement. I let him go his way, or hold him back; and so on" (PI, 208). The acts of agreement, rejection, encouragement, etc., Wittgenstein says, "will be of various kinds, and many such acts will only be possible if the pupil responds, and responds in a particular way" (BB, 89-90). In some learning processes, then, the child may anticipate encouragement or disapproval. She knows when she makes a correct association based on the teacher's reactions and gestures (PI, 208).

These observations show the grammar of expressions like "The children are learning the colors" and similar ones. In this case, the grammar of learning is closely connected with concepts like teaching, repetition, regularity, doing the same, memory, example, training, practice, skill acquisition, behavior, understanding and testing the learner's understanding, giving/receiving and obeying orders and rules, etc. As stated above, learning the colors means acquiring an ability. Thus, learning the colors presupposes a change in the student's behavior (PI, 157). Think about the importance of the body in this process. When requested, the student points to a certain color to show she understands the order. The infant utters the name of "*this* color" and she becomes able to correct herself when she blunders. She distinguishes between the colors. She becomes a user of language-games having questions like "Which is your favorite color?" and similar ones. The learning process, then, is reflected in what the student does (see Hardwick 1971, p. 87).

The overview we achieve with the description of the use of the concept of learning prevents us from trying to answer questions like "What happens inside the children's minds when they learn the colors?". The reason is that whatever answer we could offer to it (mentalist, behaviorist, Platonist, etc.) does not contribute to our understanding of the concept of learning. This kind of questions and their answers play no role in the description of the actual applications of the concept.

Before considering the next instance of human learning, let us stress that this language-game represents just one possible case of color learning. We can imagine different circumstances and different methods for children to learn this, that is, we can imagine different uses for expressions like "The student is learning the colors". Furthermore, this language-game shows merely the use of the concept of learning concerning color-words. As stated above, there are countless instances of human learning. Think, for example, how different are the circumstances in which we would say that a child is learning *ballet*. Which are the methods to teach this art

form? Which role does the body have? Consider the importance of concepts like flexibility, stability, performance, etc. How would we teach the vocabulary involved in this practice? Think about concepts that refer to certain body positions, like *arabesque*, or concepts that indicate the style of particular movements, like *allegro*.

The last case we want the reader to consider is an —extremely complicated— imaginary scenario offered by Wittgenstein in RPP I, 93-101. Imagine how colonizers taught new religions, customs, and practices to the inhabitants of a certain region. The native people have a completely different language than theirs, different practices, different culture and social organization, etc. Which could be the tools, methods, and ways to teach them a new language, to calculate the way the colonizers do or, furthermore, a new faith. The colonizers are interested in the language of the native people, in their behavior and psychological utterances, etc. to teach them, say, to pray. Which would be the criteria for the colonizers to say that these people are starting to believe in God, or the circumstances for the use of expressions like "They are learning what 'God' means", etc.? This situation represents an important, legitimate use of "learning". However, this case does not belong to the more familiar instances of human learning. To learn a new faith, a new culture, a new worldview, belongs to a less familiar and, we could say, elaborated use of our concept.

Color-words, ballet dancing, and a new faith have considerably different circumstances, teaching methods, learning processes, and are related to different forms of behavior of the people being taught. This shows us the enormous pliability of the concept of learning.

2.2.2 The language-game of autodidacticism

The last section shows how the concept of learning works in a familiar scenario. Although the language-game represents a simple and commonplace case, the use of the concept is difficult to describe. There are several presuppositions, aspects and nuances to delineate. Furthermore, the language-game and the last two examples show how people learn different things involving different circumstances, teaching methods, and intellectual and/or bodily capacities. The last example, namely, learning a new worldview, shows part of the complexity of the use of the concept. This stands for, as we called it earlier, a less familiar instance of human learning. This case, instead of representing a counterexample of the meaning of *learning*, works as a complement of its use, one that, although not very usual, is a case worth considering in, for instance, philosophy of religion. In this section we will explore another language-game of *learning*, a less familiar one. Both the familiar and less familiar language-games resemble each other according to some of their features, but they appear substantially different when we consider others. Consider the case of autodidacticism or self-teaching. This concept is applied in many practices, for instance, chess, science, painting, music, literature, football, among others.

Imagine the following case. An infant is introduced to music. Although not professional musicians, her parents play records frequently, so the child is exposed to different styles and genres at a young age. At some point they show to their daughter the sound of different instruments. The child manifests interest, in particular, for the guitar. They show her how this instrument works, that is, how one is to hold it, where to place it in the lap, and how to produce sound with it by pressing the

strings against the fret-board. Once the child is familiar enough with the basics, they teach her to play the different chords.

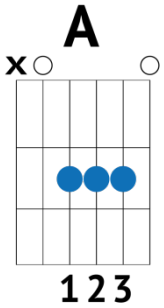


Figure 2.3: *A* Chord Diagram

"This picture indicates how to play the A chord", they say (see Figure 2.3). They show how she should place the fingers while strumming the guitar. "This is how A sounds", they say after playing the chord. They invite their daughter to repeat what they do. It is natural to suppose that at the beginning the child struggles to produce a *clean* chord. Let us imagine that they repeat this course of action for every major chord several times. After a while, the child begins to be familiar with the different chords and the ways to play each of them. By the way she reacts and follows what her parents do, they realize their daughter is skillful. Someone might say that she memorized the notes and chords quickly.[9] She does not need to receive the same instruction twice. She manifests understanding and proficiency. In such a way that is pleasant to others, the child combines chords and notes with easiness. While hearing a note, she does not need to know its name in advance to reproduce it. As musicians say, she is able to *play by ear* the same note her parents play. They encourage her to keep playing, to keep developing her skills. Gradually, they realize that the child can go on without any further instruction. She likes what she does and becomes able to interpret a given song, to play its correspondent chords, just by listening to them, without the aid of any kind of written music. The child plays frequently, sometimes at home, sometimes at school.

Imagine that one of the school professors has seen her play and knows her musical abilities. Impressed by the girl's musicianship, he talks to others: "You should see her. She is an outstanding guitar player. Moreover, she taught herself. She is an autodidact." What does this last phrase mean? How is this expression being used?

The professor wants to convey that her student has learned to play the guitar without what we call *formal instruction*. The child has not attended music lessons or any other kind of professional training. No teacher has taught her neither how to sit with the instrument nor how to hold it. All she has learned has been by copying what her parents and the musicians she has watched in videos do. Another important remark is that she has not learned any music theory. She is not familiar with concepts like texture, timbre, or harmony. As we have seen, she was introduced only to the main chords and their pictorial representations. She has little or no idea about musical notation.

[9]Frequently, the concepts of memory and learning are closely related. Sometimes we use expressions like "She memorized x" to mean that "She learned x". Both expressions are compatible. It is interesting to compare this use of the concept of memory with the analysis of this concept in *Three Lectures on Memory* (Malcolm 1963).

We could say, then, that the child is able to correct herself, that she can identify at least some of the mistakes she makes in her learning process. She knows and recognizes if a particular note or chord is the correct one or not. She might say to herself, "*This* is the way the chord sounds better!", while strumming the guitar, or "This way of playing the chorus is wrong", etc. One way of understanding this is to say that the child plays both the role of the apprentice and the role of the instructor. This is possible because the concept of guide or instructor does not refer to a particular person, but to a role one plays in a certain context.

This does not mean that the criteria for learning something in the case of autodidacticism are established by the one that learns by herself. Otherwise the concept of autodidacticism would entail some kind of mental and/or linguistic privacy. If the criteria for learning were established by the autodidact only, the concept of learning by oneself would collapse, because there would be no distinction between *learning something* and *believing to learn something*. Learning by oneself means that the learner plays, to some degree, the role of the teacher, but the criteria for learning remain public. The standards for corroborating that someone knows how to play a certain instrument do not depend on the subject only, but in the community of experts and people that compose the practice of music.

Another remarkable feature of our language-game is that the child displays better skills for music than some people do. Someone might say, for instance, that she has a sharper ear than most of music students. Or she might have what some musicians call *absolute pitch*, namely, "the ability to produce a note of a given pitch on the absence of a reference note" (Deutsch 2013, p. 141).[10] However, it is worth noting that, although impressive, this feature of our language-game is not crucial. It is possible to imagine situations of autodidacticism in music (and in other fields) in which these special skills do not play any role. A trained ear can be —and most of the time is— developed. Someone can start learning music by herself without having absolute pitch. Likewise, it is imaginable that someone who has this special skill is not interested in music. Thus, the concept of autodidact and the concept of exceptionally gifted can be related but are not interchangeable.

As an autodidact, the child has, we may say, the power to decide about her learning process. After her parents introduced her to the guitar, nobody compels her to learn music. She decides to learn, and the reasons she may have can be manifold. In contrast with the first language-game, her training is not decided by a third person. She decides what to learn, how to learn it, and at what pace or rhythm. She plays music she finds pleasant, she learns from playing records, watching videos, rehearsing with friends, from books, etc. she may spend days or even months on a song she wants to perfect or on a composition she is working on. This last point is connected with an underlying idea about the child's personality. In general, concepts like willpower, tenacity, motivation, and discipline are important elements for understanding the concept of autodidacticism. There are disciplined and tenacious people that are not autodidacts. But it is difficult to imagine someone that is an autodidact in a given field and is not disciplined, motivated, or tenacious concerning such field. Additionally, the autodidact relishes the practice in question, in this case playing music. Just as above, non-autodidact people can enjoy what they learn. But it is problematic to imagine a case in which the autodidact is learning

[10]This topic has been studied in many fields, such as psychology, neuroscience, and genetics. It deserves further philosophical consideration.

something she dislikes.

Not having formal instruction, deciding about one's learning process and the concepts of motivation, tenacity, and willpower are some of the elements that make this a less familiar case of learning. Once more, the child can interpret various compositions proficiently. She is also able to produce agreeable music. In most of the cases these skills are developed after long periods. People practice for years to accomplish this. What the child can do with the guitar is simply impressive. When thinking about her case, it is natural to focus on what is most striking, namely, how young she is, the skills she has been able to master, and the fact that she lacks theory and professional instruction. Because of the rareness of this situation, it is easy to neglect simpler and more familiar aspects, all of which lie before our eyes (PI, 129). The description of these elements, however, is of the utmost importance, for it contributes to the understanding of the child's case. Let us ask, which are the components that both language-games share? What lies in their background? Let us explore some of their similarities.

Both instances of human learning are related to the concept of reacting to the instructions and orders of someone else. In the first stages of different learning processes we imitate and copy forms of behavior, movements, and actions of the people that guide us (see RPP I, 163). When introduced to the guitar, our musician copied and imitated what her parents did with the instrument. We can imagine the parents saying something like "You hold the guitar like *this*", exemplifying how the child is supposed to act. We also can imagine that the parents placed the instrument correctly in the child's lap while saying "Like *this*!" setting the child's hands in the right position and so on. Likewise, in the language-game of learning the colors, the teacher invites the children to imitate her, saying something like "Repeat after me! *This* is red" while pointing to the color in question. In both situations, the learning process begins with the children mirroring the parents' and the teacher's behavior, respectively. Consider the following idea.

> the origin and the primitive form of the language-game is a *reaction*; only from this can more complicated forms grow. (CE, 395. My emphasis)[11]

In both language-games, then, children's reactions belong to the ways human beings learn.[12] Paraphrasing Wittgenstein, we could say that the ways in which we learn —including different kinds of reactions, behavior, imitation, encouragement, etc.— belong to our natural history just as "walking, eating, drinking and playing" (PI, 25).

Another important point in common of both language-games is that at some point, with enough training, the learner goes on by her own. By having seen what her parents do and by having been taught how to act, the child starts to play the

[11] Also consider: "it is characteristic of our language that the foundation on which it grows [in our case the reactions children have] consists in steady ways of living, regular ways of acting" (CE, 397). These notes are about the concept of cause. However, they can be used to show how the concept of learning is used in different circumstances. In relation to this point Meredith Williams suggests that "Wittgenstein is struck [...] by the extent to which we are like his builders (despite their unrealistically impoverished language game) rather than logicians or scientists. The complex and sophisticated normative structure of the language games of the logician and scientists rests on a bedrock of shared primitive techniques acquired through training" (Williams 1999b, p. 6).

[12] See Hertzberg 2011 for the ongoing discussion about *facts of nature*, *primitive reactions* and *concept formation*.

chords. She follows the instructions given by her parents. She applies what her parents say to her. At the beginning she may not produce clean sounding chords, but this is expected to come with training. Similarly, once the teacher thinks (based on certain evidence) her students are familiar enough with the colors and the color-words, she may test the children's understanding by asking questions like "What color is *this*?" while pointing to a certain color. Just as in the other case, they may err. At a certain point in the training, the children are said to be able to "go on by their own". And this means that they apply the color-words correctly. If children proceed accordingly to the given instructions, it is natural to imagine that expressions of agreement and approval take place to encourage the children to "do the same". Let us remember that in the rule-following discussion Wittgenstein stresses the importance of the fact that it is not possible to follow a rule just once. The following remarks show that our practices, games, and institutions presuppose regularity and are grammatically related to the concept of repetition and of doing the same.

> The application of the concept "following a rule" presupposes a custom. Hence it would be nonsense to say: just once in the history of the world someone followed a rule (or a signpost; played a game, uttered a sentence, or understood one; and so on). (RFM VI, 21)
>
> Following a rule is a human activity. (RFM VI, 29)
>
> A game, a language, a rule is an institution. (RFM VI, 32)

These traits of the use of the concept of learning are related to the concepts of guidance, training, practice, and repetition. In other words, in these language-games children do not learn by chance. Our young musician spends hours practicing the different chords and trying to learn different songs. Similarly, children need time and practice to become able to apply the color-words correctly. In both cases, children are said to follow certain courses of action several times. As these examples show, the methods and ways of teaching are different depending on the learning process in question.

Furthermore, in both language-games "learning" is understood as *ability acquisition*. In the second case, the child is able to do something, namely, to produce articulate sounds with a particular instrument. Based on training, practice, and repetition, she can perform or interpret a composition she likes. This presupposes a change in her behavior (PI, 157). In contrast with people that do not know how to play, she behaves differently, her actions while playing the instrument are, we may say, coherent. Likewise, when learning the colors, children are said to acquire various skills. As Hertzberg suggests, the acquisition of color-concepts is not to be understood only as the "problem of learning to classify things [and colors] correctly" (Hertzberg 2011, p. 360). This means that children do not only learn and become able to classify, identify and recognize objects by their color. Instead, children learn the use of color-words in a broad sense, that is, they become users of the complex vocabulary related to colors (see RPP I, 142). The abilities children gradually acquire allow them to be part of our practices, for instance, being able to say which color they like the most. This practice is connected with language-games where inclinations, preferences and choices are important elements. Likewise, they become able to combine garments, to solve puzzles like the Rubik's cube, to play games

where colors are important like cards or billiard, to drive and recognize street signs and traffic lights, to understand caution and danger signs in a laboratory, etc. As we can see, the familiar aspects of the use of the concept of autodidacticism are also important. Without these traits, the concept of self-teaching and the concept of learning, in general, would not have their actual uses.

The last point we want to consider is that although the concept of experience does not seem to be necessary to describe satisfactorily neither of both instances of human learning, the concept of experience is related to the case of autodidacticism in a higher degree than to the case of color learning. Imagine someone says: "Children learned the colors through experience". This expression would invite to ask further questions, for instance, "What do you mean by experience? In which sense are you using this concept?" Using "experience" in this case seems to exceed what is required for describing how children learn the colors. If this person means that "Children need to see the colors to learn them", we would reply that learning the colors presupposes seeing them, and this is a proposition of the grammar of the expression "learning the colors". Similarly, "Children have experience in distinguishing the colors" is equivalent to say "Children (can) distinguish the colors". Thus, the concept of experience does not seem necessary in this case. In contrast, the concept of experience seems to be somehow related to the case of autodidacticism in the following sense. "Learning by experience", in some cases, means that there are abilities that are acquired by actually exercising them. In other words, abilities "we learn by doing; for example, we become builders by building and lyre-players by playing the lyre" (Aristotle 2004, 1103a-b). "Learning to play guitar" represents one of these cases. It is not possible to learn to play a given piece of music without in fact rehearsing it.

To sum up. By describing both instances of human learning we can distinguish more accurately the relations this concept holds with others. The grammatical network of expressions like "She learned the colors" and "She learned to play by herself" include various notions, for instance, reactions and responses, a change in the learner's behavior, regularity in the learner's actions ("doing the same"), memorizing, training, the criteria and evidence that allows us to say that the child mastered a technique, (gradual) ability acquisition, examples as right courses of action, guides or authority figures (in a higher degree in the case of the first language-game), etc. The idea of family resemblance is key to understand that the concept of learning is applied to manifold cases, not because they share a single feature that legitimizes the use of this concept, but because of the overlapping similarities among them (see Glock 1996, p. 121). In other words, none of the features these cases share represents the essence of human learning, and thus, neither of them guarantees by itself that we have an instance of learning. For example, not all the abilities we acquire are learned. Blinking, crying and breathing are abilities we develop when we are born, and we would not say we learn them. Likewise, we can make use of examples to guide our actions in a certain activity without learning it, for instance, when following the instructions for solving a certain puzzle or assembling a certain piece of furniture just once, without memorizing the procedure.[13]

[13]Meredith Williams' controversial but reasonable contention about the concept of learning, mainly in mathematics, is that the process of learning (how we learn) plays a constitutive role in the activity we learn (what we learn). According to this author, the relationship between the way of learning and its product is not contingent. The concept of competent user, say, in mathematical

operations, presupposes training in such techniques. She asserts: "Training in techniques creates the regularities of behavior necessary for any judgment of sameness, and it is for this reason that one can say the process of learning is constitutive of what is learned. Moreover, this explains why the psychological sense of the obvious is grounded in the grammatical, not the other way around" (Williams 1999a, p. 215). Her position needs to be contrasted not just with the acquisition of mathematical concepts and rules, but with more instances of human learning.

Chapter 3

The Use of *Learning* in Machine Learning

The objective of this chapter is to describe the technical usages of the concept of learning in machine learning. The aim is to show how the concept of learning is applied to machines and software. Machine learning in general is an extremely complex and continually growing discipline. Its methods and techniques are improved constantly. Moreover, the scope of problems this field addresses is broadened every day. Offering an account of the application of the concept of learning in all machine learning models surpasses the spectrum of this investigation.

Wittgenstein proposes to analyze simple language-games and "primitive kinds of use" of words to clarify more complicated concepts and forms of language (PI, 5). Likewise, in this chapter we are not going to describe the functioning of the contemporary pinnacles of machine learning algorithms, such as Google Translate in natural language processing for example (Turovsky 2016). Instead, we will focus on and analyze the use of the concept of learning in two basic machine learning models, each of which represents a supervised and an unsupervised approach, respectively. These approaches have been chosen for the reason that their names suggest a relationship with the previously analyzed instances human learning, that is, the familiar case which is connected to the concept of supervision or instruction, and the case of autodidacticism, the less familiar instance of human learning which is not connected, or connected in a less rigid sense, to the concept of supervision or instruction. The displayed uses of the concept of learning in these models remain relevant for at least some of the more complex cases of machine learning in which more elaborated data, programming tools, and mathematical models are employed.[1]

To understand the technical usage of *learning* in machine learning, this chapter is divided into three sections. First we explain some of the basic key concepts in machine learning using a simple and typical example. After, we introduce both machine learning models and show how they work. We offer a succinct explanation of the code and the mathematics that lie behind them. In the last part the use of the concept of learning in both models is analyzed and described. This chapter contains technical language from machine learning. However, it is necessary to understand how the models work to gain clarity and prevent incomprehension about the application of mental concepts to machines and software.

[1] After going through this chapter, to see how deep learning and neural networks work see Zhang et al. 2016.

3.1 Basic terminology and general procedure

Machine learning, in general, is composed of two wide fields, mathematics and computer science. Accordingly, machine learning encompasses multiple practices, such as computer and mathematical programming, algorithm studies, statistics, algebra, among many others. Because of the inherent complexity of machine learning, it is necessary to offer some definitions and terminology to understand what the models in the next section do and how they do it.

First of all, the difference between a machine learning model and a traditional computer program is that in the latter the programmer implements a sequence of defined instructions to be followed to transform given input into a certain output. In other words, the programmer executes an algorithm to solve a particular task. For example, the problem of sorting the names of a group of people alphabetically can be solved by one or various algorithms. A machine learning model, in contrast, is designed to solve much more complex problems, such as the prediction of human behavior, natural language processing, speech recognition, or the detection of patterns of certain diseases, such as cancer. In these cases, the model *learns* an algorithm to give an approximate solution to the problem in question (see Alpaydin 2014, p. 2).

A machine learning model *learns form experience*. More technically, Mitchell states that this kind of model is said to be "any computer program that improves its performance at some task through experience" (Mitchell 1997, p. 2). Paraphrasing this author, a contemporary example of a machine learning model is AlphaGo, a machine learning-based computer program that plays the game of Go (Silver et al. 2017). This model obtains experience by training and playing against itself. The data the model collects represents the experience that might improve the model's performance at Go while playing against other contenders. If the model's performance does improve, computer scientists say that the model learns from experience.

The most important features in machine learning are the data and the mathematical model that helps us to handle the data. Let us explain this with an example. Think of a program that filters spam emails from authentic ones. Because an email, spam or not, is only a "file of characters", there is no sequence of defined rules that determines exactly what constitutes one or the other. However, "[w]hat we lack in knowledge, we make up for in data" (Alpaydin 2014, p. 2). In online libraries and repositories thousands of examples of spam and non-spam emails are available. A crucial step is ordering and preparing the data correctly. Among other things, this means that the data must not have incorrect or missing values. According to computer scientists the quality and quantity of the data set have a direct influence on the predictions of our model (Alpaydin 2014, p. 1ff). In our case, the data is ordered and labeled in two classes, spam and non-spam emails. Then the machine reads this data, and the programmer trains the model based on these examples. The machine will detect patterns and regularities in the data, for instance, typical expressions and words of a spam email. As a result, the machine will be able to *infer* from previous experience the label or class of an email the machine has not seen before.

The inference or generalization the machine can make depends on a mathematical model, and more precisely, on a statistical one. Certain kinds of mathematical models work better for specific problems and data. For example, one of the solutions for determining whether an email is spam does not consist of an integer or number,

but of a yes/no output (spam = yes; spam = no). These are called *classification* problems. The Support Vector Machine (SVM) is one of the models computer scientists use to deal with these kinds of data and problems, i.e. when the distinction between different classes of items, expecting to obtain a binary output, is needed.

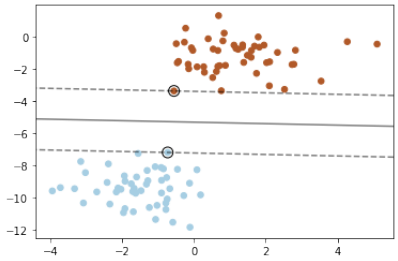


Figure 3.1: Support Vector Machine Graph

Figure 3.1 is the graphic representation of the output given by an SVM model. This algorithm is used for establishing the relation between two or more data points by implementing a classification analysis. The computer reads the data previously labeled by the programmer. The programmer trains the model until the necessary patterns are detected. And the programmer implements a classification analysis in the model, so the machine can infer and predict the class of a new item. The outcome of the analysis is the line that divides the data into two classes. Thus, based on "example data or past experience" the model learns, that is, the performance of the model is optimized (see Alpaydin 2014, p. 3).

The data and the mathematical model, thus, are fundamental in machine learning. Both aspects are closely related. Without the mathematical model computer scientists could not analyze the data or it would be much more difficult for them to get the expected results and inferences. Likewise, based on well-ordered data, its features, complexity, and amount, a mathematical model is selected to have a better understanding of a given analysandum.

3.2 Two machine learning models

Machine learning represents an enormous number of techniques and practices. Its models can be categorized by the problems they address, the kind of data they process, and the type of results they offer. Depending on these factors, there are several categories of machine learning problems, for instance, supervised, unsupervised, semi-supervised, and reinforcement learning. In turn, each of these approaches comprises many different models, for example, neural networks, decision trees, probabilistic classifiers, regression analysis, and clustering (see Alpaydin 2014, Introduction). In this chapter we will focus on the last two of them. In general, the different machine learning models can be divided by the task they perform, that is, classification (and) or prediction. The model above, Support Vector Machines, is an example of a supervised learning model that performs classification analysis.

Certainly, the scope of this investigation does not include offering an account of each of these models. Furthermore, this would be irrelevant for our purposes. The aim, instead, is to briefly describe the functioning of two basic machine learning models to understand how computer scientists apply mental concepts to software.

Our concern in particular is with the concept of learning. Moreover, the explanation of these models, although simple, remains relevant for understanding, at least in principle, how more complex models work.

The following are, respectively, examples of supervised and unsupervised machine learning models. In a few words: to work correctly, the first one needs training data, that is, data that has been previously labeled, so the model can associate correctly future data with the right features. In unsupervised learning, the expert does not provide the machine with previously labeled examples. These programs, based on a mathematical model, can detect patterns and structures in the data automatically, that is, with little or no help from the expert. The corresponding names of these models are related to the concepts explored in the last chapter, that is, the familiar instance of human learning, where a supervisor or teacher guides and trains the student in a particular activity, and the less familiar instance of human learning, in which the learner does not need, or needs in a slighter way, the aid provided by the expert.

The models are implemented in Python. Unfortunately, for inexperienced users —very much like the author of these pages— this programming language, as many others, can be cryptic. Programming languages like Python are instances of formal languages of high complexity. For this reason, and for the sake of clarity, (i) the models are explained through concrete imaginary situations, and (ii) we explain the code in so far as it is necessary for the inexpert reader to understand what happens in the models' processes. For each block of code we introduce the sign # followed by a corresponding explanation. However, in the Appendix A the reader can find a comprehensive list of the repositories and tools, along with their sources, from which we obtain the data, code, and graphics to build up the machine learning models.

3.2.1 Supervised learning: linear regression

Basically there are two types of supervised learning models, classification and regression. The models in the first group offer a class as outcome, that is, a binary output, like "yes/no", "spam/not spam" —as our example above— or "positive/negative" in a medical diagnosis. The outputs of the models in the second group, in contrast, are values, real numbers. In general, the data has some kind of correlation that the model can determine. Imagine the following situation.

We want to build a machine learning model that can infer the future performance of a group of students, that is, their final grades, based on academic and (some) personal information. We can accomplish this using a linear regression model. We start by importing the necessary tools for solving this problem (see Figure 3.2).

```
# The user must install the following packages, libraries and tools
# in the programm. By doing this, the machine is able to read the data
# and to process it using a mathematical model, in this case, linear
# regression.

import tensorflow
import keras
import pandas as pd
import numpy as np
import sklearn
from sklearn import linear_model
from sklearn.utils import shuffle
```

Figure 3.2: Linear Regression Code 1

The high school database contains the following information for every student: sex, address, family size, the reason the student had to choose this school, study time per week, extra-curricular activities, the last math grades, etc. The reason for choosing a linear regression model is the strong correlation between some of the features of the student database, for instance, the number of school absences and the grade of a certain student.

Briefly, linear regression is the statistical approach for modeling the correlation between an independent variable (x), and a dependent variable (y), through the best-fitting line. This means that based on x, the variable representing the value we already know, we can predict the unknown value for the dependent variable, y. Figure 3.3 represents a linear regression model, where we can see the best-fitting line and the correlation between x and y: when the value of x increases, so does the value of y.

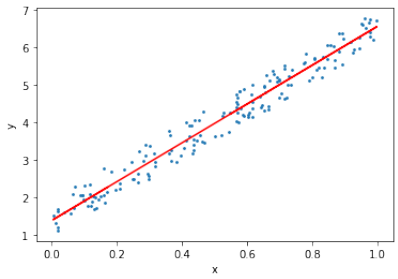


Figure 3.3: Linear Regression Graph

Going back to our case, we upload the database of the whole class to the model (see Figure 3.4). We order the machine to read and print the data (see Figure 3.5 and Figure 3.6).

```
# This line of code allows us to import the data from the computer to our
# colab document. In this case, we import the document that contains the
# students' data.

from google.colab import files
uploaded = files.upload()
```

Choose Files | student-mat.csv

- **student-mat.csv**(text/csv) - 56993 bytes, last modified: 4/12/2012 - 100% done

```
Saving student-mat.csv to student-mat.csv
```

Figure 3.4: Linear Regression Code 2

```
# Once the file is imported to our colab document, this command is used to
# order the model to read the data.

data = pd.read_csv("student-mat.csv", sep=";")
```

Figure 3.5: Linear Regression Code 3

```
# This tool helps us to visualize the data the model is working with.
# The first column represents each student. In the first row we can see all the
# data that has been collected for every student.

print(data)
```

```
     school sex  age address famsize Pstatus  ...  Walc  health absences  G1  G2  G3
0        GP   F   18       U     GT3       A  ...     1       3        6   5   6   6
1        GP   F   17       U     GT3       T  ...     1       3        4   5   5   6
2        GP   F   15       U     LE3       T  ...     3       3       10   7   8  10
3        GP   F   15       U     GT3       T  ...     1       5        2  15  14  15
4        GP   F   16       U     GT3       T  ...     2       5        4   6  10  10
..      ...  ..  ...     ...     ...     ...  ...   ...     ...      ...  ..  ..  ..
390      MS   M   20       U     LE3       A  ...     5       4       11   9   9   9
391      MS   M   17       U     LE3       T  ...     4       2        3  14  16  16
392      MS   M   21       R     GT3       T  ...     3       3        3  10   8   7
393      MS   M   18       R     LE3       T  ...     4       5        0  11  12  10
394      MS   M   19       U     LE3       T  ...     3       5        5   8   9   9

[395 rows x 33 columns]
```

Figure 3.6: Linear Regression Code 4

```
# This command prints just the five first rows of the data file.
# In our case, the five first students along with their corresponding features.

print(data.head())
```

```
  school sex  age address famsize Pstatus  ...  Walc  health absences  G1  G2  G3
0     GP   F   18       U     GT3       A  ...     1       3        6   5   6   6
1     GP   F   17       U     GT3       T  ...     1       3        4   5   5   6
2     GP   F   15       U     LE3       T  ...     3       3       10   7   8  10
3     GP   F   15       U     GT3       T  ...     1       5        2  15  14  15
4     GP   F   16       U     GT3       T  ...     2       5        4   6  10  10

[5 rows x 33 columns]
```

Figure 3.7: Linear Regression Code 5

We also realize that some of the information may not be important. For our purpose, that is, determining the future performance of the students, we must filter out the irrelevant portion of the data. We may see that there is no correlation between the grades of the students with, for instance, the reason they had to choose this school or the student's age. We need to focus on the features that seem to correlate with the grades. The data has to be prepared in the right form. If this is not the case the computer will not offer accurate results. In this case, we will take seven attributes: first, second and third grade, study time per week, failures, absences, and free time of each student (see Figure 3.8).

```
# With this line of code we are able to select the features we want to train
# our model with. It may happen that some of the collected data is irrelevant.

data = data[["G1", "G2", "G3", "studytime", "failures", "absences", "freetime"]]
```

Figure 3.8: Linear Regression Code 6

These attributes will help to know the future performance of each student. More concisely, the machine will read the data, and with the help of a linear regression model, the machine will be able to *predict* or *infer* their final grades. In Figure 3.9 the features we want to work with are represented by means of a table.

```
# On the basis of having specified the relevant data, we order the machine
# to print the new set of data.

print(data.head())
```

```
   G1  G2  G3  studytime  failures  absences  freetime
0   5   6   6          2         0         6         3
1   5   5   6          2         0         4         3
2   7   8  10          2         3        10         3
3  15  14  15          3         0         2         2
4   6  10  10          2         0         4         3
```

Figure 3.9: Linear Regression Code 7

Once we have the data in the correct form, we order the machine to predict the label "G3", which stands for the final grade of each student (see Figure 3.10). A label is what the model predicts, based on the specified features, in this case, "G1", "G2", "studytime", "failures", "absences" and "freetime".

```
# This line of code orders the machine to predict the value for "G3" for
# every student.

predict = "G3"
```

Figure 3.10: Linear Regression Code 8

After this, we need to adjust our data into two arrays. The first one, represented by x, is composed of the attributes of each student. The second one, y, represents the label we want to predict, in this case, "G3" or the final grade of each student (see Figure 3.11).

```
# Lines of code for setting up the data into two arrays, the first of which
# represents our attributes, or the known data. The second array represents
# the label we want to predict, in this case, "G3". In the first line of code
# we order the machine to separate (or "drop") "G3" from the attributes. In the
# second line of code, we order the machine that "G3" is to be considered the
# label we want to predict.

x = np.array(data.drop([predict], 1))
y = np.array(data[predict])
```

Figure 3.11: Linear Regression Code 9

Once both arrays are established, we need to separate both of them into two different sets: x and y training data, and x and y testing data (see Figure 3.12). In this case we will train our model with 90% of the samples. The other 10% of the data will be used to test whether the machine accurately learned the patterns of the data.

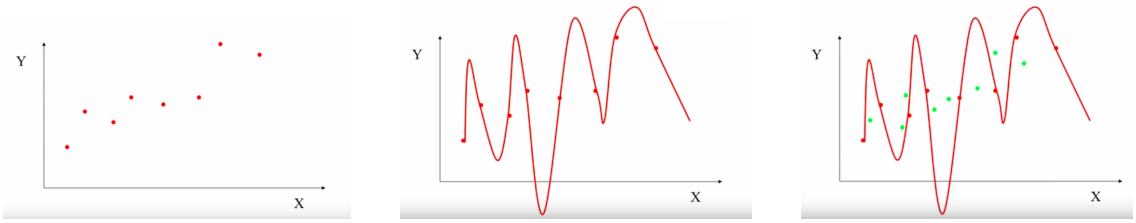
```
# Line of code that separates each array into two different variables: training
# data and testing data. This code also specifies which amount of data will be
# used for testing our model, in this case, 10%.

x_train, x_test, y_train, y_test = sklearn.model_selection.train_test_split(x, y,
                test_size=0.1)
```

Figure 3.12: Linear Regression Code 10

The more training data we have, the more the model can learn. However, it is crucial to divide our data into training and testing arrays, respectively. Otherwise, if we decide to train our model with the data we are going to test it, the model could not estimate the generalization performance on future data. This may lead the model to memorize the patterns of the data set, and when tested, instead of performing the inference or prediction, the model may just repeat the patterns previously memorized. In other words, the training process would be perfect, the machine will offer the exact pattern for the training data, but this does not guarantee that the model will perform accurately with unseen data in the future. This problem is known as *overfitting*, which can be explained as memorizing without learning.[2]

In Figure 3.13a we can visualize some data. If the model is trained with all the data points available, one of the possible patterns the model can give is represented in Figure 3.13b. The data points in Figure 3.13b are correctly explained by the pattern given by the model. However, when presented with new, unseen data, the model may not give an account of its actual pattern, as displayed in Figure 3.13c. Testing a model does not mean testing it with the training data, with data that the model has memorized before. Testing a model means assessing its performance concerning new, unseen data. For this reason it is crucial to split the data into training and testing sets.



(a) (b) (c)

Figure 3.13: Overfitting

Back to our example, once we have divided the training data from the testing data, the learning process begins. In this stage of the process we train our model, that is, we order the machine to determine the best-fitting line for our training data using the linear regression model (see Figure 3.14).

[2]The opposite problem is *underfitting*, in which the model does not manage to capture the structure of the training data. I will not expand on this.

```
# The first line of code specifies that we are implementing a linear
# regression model on the training data. The second is used to produce the
# best-fitting line for the training data, which represents the
# 90% of all of our data.

linear = linear_model.LinearRegression()
linear.fit(x_train, y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

Figure 3.14: Linear Regression Code 11

The machine learned the pattern of the data. In other words the machine determined the best-fitting line for the specified training data. The next stage of the process is to test our model with the data we saved for evaluating it. The assessment will determine whether the machine successfully learned the patterns of the data. The accuracy of the model is 86% (see Figure 3.15).

```
# Line of code that gives you the accuracy of the best-fitting line our model
# determined previously on the basis of our testing data. We test whether the
# model learned the patterns of the data or not. The accuracy is printed with
# the help of the second line of code.

acc = linear.score(x_test, y_test)
print(acc)
```

```
0.8693895473381748
```

Figure 3.15: Linear Regression Code 12

The model is not 100% accurate. Perhaps the model can offer better results if we train it with more relevant features or with more data in general. However, our objective is to have an approximation about the future performance of a group of students. Human behavior is complex and not all of its aspects can be quantified or measured.[3] For this particular problem, thus, we can say that the model works. For other problems, such as determining whether a certain patient displays patterns of a certain disease, this same accuracy may be considered low.

The final step is to use our model with the testing data. In other words, we will use our model to predict or infer the final grades based on each student's attributes. In Figure 3.16 the first column stands for the model's predictions of each student's grades. The second column represents the features of each student. The third column is the grade the student got. For the first student, the machine predicted a grade of 14.78. She had 15 in the first two grades, 2 in study time, zero failures and absences, and 1 of free time. The final grade of this student was 15. Row number eleven represents a mistake of our model. The machine predicted 6 as the final grade. But the actual grade the student got was zero.

[3]We can imagine a situation in which a student displays poor academic performance the whole semester, but she gets an outstanding grade for the last exam. This is not contradictory. She may have received motivation, or she may have suddenly understood the subject in question, etc.

```
# Line of code that orders the machine to predict the final grades of each
# student with the testing data, that is, with the data the model has not
# seen before.

predictions = linear.predict(x_test)

for x in range(len(predictions)):
  print(predictions[x], x_test[x], y_test[x])
```

```
14.78047311443554 [15 15  2  0  0  1] 15
14.930424142976403 [14 15  2  0  4  2] 15
12.482217069191991 [14 12  1  0  4  4] 11
11.823077887069743 [11 12  1  0  2  3] 11
7.7771251964369075 [6 9 1 1 4 2] 8
13.957617807024684 [15 14  3  0  6  2] 14
10.566793423193603 [11 11  2  0  0  3] 10
17.365758012034537 [16 17  2  0  0  4] 17
9.2294494264479 67 [ 9 10  3  0  4  3] 10
5.584432289751737 [7 6 1 0 5 4] 7
6.076742693235172 [7 7 2 1 0 5] 0
9.39104617028985 [10 10  3  0  0  4] 9
11.99398604198552 [12 12  2  1 12  4] 13
4.71623181937317 [8 6 2 2 2 3] 5
6.592860195920759 [ 9  7  2  1 20  2] 8
12.347919075758533 [11 13  4  0  6  3] 14
10.002442813136275 [12 10  2  0  8  3] 11
9.876022802960644 [11 10  2  1 12  4] 10
9.715868471324423 [12 10  3  0 10  2] 12
7.298056361835869 [9 8 4 0 2 5] 8
11.677243417482524 [13 12  3  0  1  3] 12
7.717529797501019 [ 9  8  2  1 15  4] 8
12.823584297355435 [12 13  2  0  4  3] 13
9.429663699636922 [ 9 10  2  0  4  3] 10
```

Figure 3.16: Linear Regression Code 13

3.2.2 Unsupervised learning: clustering

In contrast with supervised learning, in unsupervised learning we do not provide the machine with labeled examples that guide the correct procedure. We only provide the machine with input data. The objective of this kind of models is not to predict or infer an output from given information. Instead, the goal is to gain some insight into the present behavior of a certain set of unlabeled data by finding and learning —based on a mathematical model— its patterns and structure (see Alpaydin 2014, p. 11).

There are different clustering techniques. In this investigation we will focus on K-Means clustering. This model analyzes a certain data set and divides it into clusters or subgroups in such a way that the data points of one cluster are similar among them, but different from the data points of another cluster. Clusters are defined by centroids or center points, "which are the typical representatives of the groups" (Alpaydin 2014, p. 167). The number of clusters is usually specified beforehand (see Alpaydin 2014, p. 162).

In relation to Figure 3.17 (Bishop 2006, p. 426) it is specified in advance that the machine must search for two clusters (K = 2). The machine learns the structure of this data set by applying a mathematical model, in this case Lloyd's algorithm. (a) A certain data set is given and both centroids have been chosen randomly. (b) The machine allocates every data point to each cluster based on its proximity. (c) The new centroids' position is defined by determining the mean average of the data points of each of the clusters in (b). The model gradually improves by repeating this procedure until there is no variation in the clusters (i). The model is not evaluated in the same sense as the previous one. That is, we cannot test the accuracy of the model from within. Nevertheless, the final output of these models must be evaluated by the expert at the end of the process. This task is more difficult because there is not only one correct result. K-means clustering models group data into k clusters regardless of whether there is such amount of clusters. Thus, the evaluation of the model

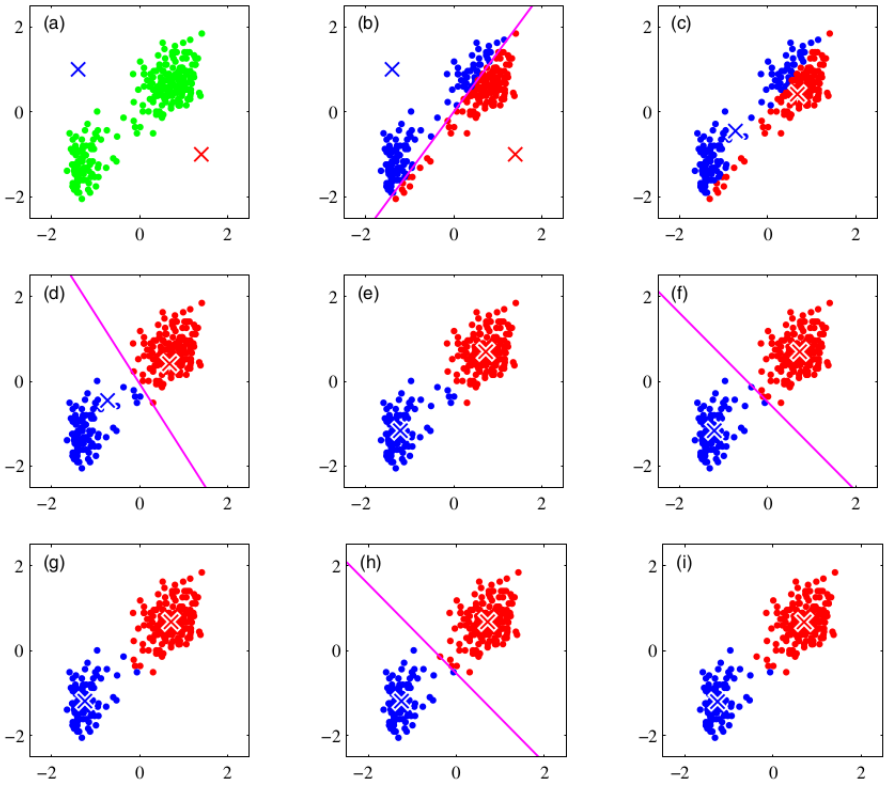


Figure 3.17: Clustering Model

does not belong to the machine learning process itself. For this reason, the external evaluation the expert makes is not going to be considered in this investigation.

Let us imagine the following situation. We want to have a better understanding of a certain group of footballers to promote a maximum wage law in the sport. In this situation, we are interested in the footballers' age and income per year. There is no strict correlation among the data. To gain the knowledge we want, we can build a clustering model. We start by importing the necessary tools and the footballers' data set to our model (see Figure 3.18). After this, we print the data set both in its original form and in a scatter plot (see Figure 3.19 and Figure 3.20).

```
from sklearn.cluster import KMeans
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from matplotlib import pyplot as plt
%matplotlib inline

from google.colab import files
uploaded = files.upload()
```

```
Choose Files | footballer_i...me_year.csv
• footballer_income_year.csv(text/csv) - 532 bytes, last modified: 2/24/2020 - 100% done
Saving footballer_income_year.csv to footballer_income_year.csv
```

Figure 3.18: Clustering Code 1

Once we have seen the data set graphically, we need to take an intermediate step. Our data values are not scaled properly. They differ significantly between them. The income axis goes from 40,000 to 160,000; while the age axis goes from 0 to 50. It is necessary to standardize the data values so the machine can give the

```
df = pd.read_csv("footballer_income_year.csv")
df.head()
```

	Footballer	Age	Income
0	1	27	70000
1	2	29	90000
2	3	29	61000
3	4	28	60000
4	5	42	150000

Figure 3.19: Clustering Code 2

```
# Line of code that represents the data in a scatter plot.

plt.scatter(df['Age'],df['Income'])
plt.xlabel('Age')
plt.ylabel('Income')
```

```
Text(0, 0.5, 'Income')
```

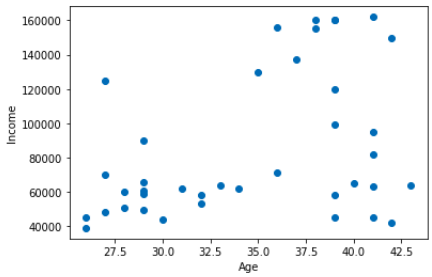


Figure 3.20: Clustering Code 3

correct results. This step, then, may not be necessary in cases where the values are scaled in advance. To see the values correctly scaled, we print the table again (see Figure 3.21).

```
# This line of code helps to scale our values from 0 to 1 in both axis.

scaler = MinMaxScaler()
scaler.fit(df[['Income']])
df[['Income']] = scaler.transform(df[['Income']])

scaler.fit(df[['Age']])
df[['Age']] = scaler.transform(df[['Age']])

df.head()
```

	Footballer	Age	Income
0	1	0.058824	0.252033
1	2	0.176471	0.414634
2	3	0.176471	0.178862
3	4	0.117647	0.170732
4	5	0.941176	0.902439

Figure 3.21: Clustering Code 4

Now the data is correctly scaled and the machine learning process begins. The estimated number of clusters in our data set is specified, in this case three (upper right corner, lower left corner, and lower right corner).[4] After, we implement the

[4]With more complicated data sets, it may be difficult to estimate the number of clusters. However, there are techniques for choosing the value of K, such as the elbow method. See Bishop 2006, p. 427.

required mathematical model, namely, k-means algorithm (see Figure 3.22).

```
# Line of code that executes the k-means algorithm in our model. The number
# of clusters is specified. The machine will determine the cluster of each
# data point.

km = KMeans(n_clusters=3)
y_predicted = km.fit_predict(df[['Age', 'Income']])
y_predicted
```

```
array([1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 0, 2, 2,
       1, 2, 1, 2, 2, 2, 1, 2, 1, 1, 0, 0, 1, 1, 1, 1, 2], dtype=int32)
```

Figure 3.22: Clustering Code 5

The machine learned the structure of the data, and predicted which of the three clusters each data point belongs to (see Figure 3.23). The names of the clusters are 0, 1, and 2, respectively. The last step is to visualize the output, that is, the clusters the machine previously determined, and to specify the location of the clusters' centroids (see Figure 3.24).

```
# Line of code that orders the machine to print the data points and their
# corresponding clusters in the form of a table.

df['cluster'] = y_predicted
df.head()
```

	Footballer	Age	Income	cluster
0	1	0.058824	0.252033	1
1	2	0.176471	0.414634	1
2	3	0.176471	0.178862	1
3	4	0.117647	0.170732	1
4	5	0.941176	0.902439	0

Figure 3.23: Clustering Code 6

```
# Line of code that specifies where are the clusters' corresponding centroids.

km.cluster_centers_
```

```
array([[0.18954248, 0.18270099],
       [0.72941176, 0.89430894],
       [0.8342246 , 0.22172949]])
```

Figure 3.24: Clustering Code 7

In the scatter plot of Figure 3.25 we can see each cluster represented by a different color. As well, the centroids of each cluster are marked with black symbols. In this case, our model worked only with input data. In the learning process there was no inference or prediction. In this sense, we can say that clustering can be understood as a descriptive method, which aim is to represent a certain data set perspicuously. However, clustering can also be understood as a predictive model —in a less stronger sense than supervised algorithms—, i.e. if we input new data points, the model will automatically place them in the corresponding clusters previously determined, based on each data point's features. With this classification, the expert counts with a better understanding of the data set, and is now able to name each cluster according to the similarities of its corresponding data points. In other words, the footballers can be classified based on their attributes represented in the plot (see Alpaydin 2014, p. 173).

```
# Lines of code that represent our data set in the form of different clusters,
# by specifying their corresponding centroids.

df1 = df[df.cluster==0]
df2 = df[df.cluster==1]
df3 = df[df.cluster==2]

plt.scatter(df1.Age,df1['Income'],color='red')
plt.scatter(df2.Age,df2['Income'],color='blue')
plt.scatter(df3.Age,df3['Income'],color='yellow')

plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color='black',
            marker='*',label='centroid')

plt.xlabel('Age')
plt.ylabel('Income')
```

```
Text(0, 0.5, 'Income')
```

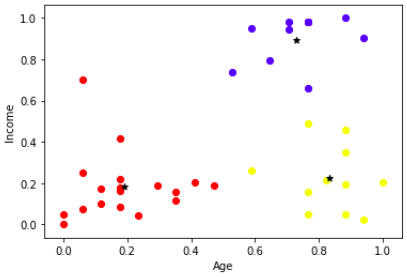


Figure 3.25: Clustering Code 8

Let me once again stress the importance of ordering the data set before the model analyzes it. Although these models are impressive, their functioning depends on correctly ordered and prepared data. If the data is not properly arranged, the results the model offers will, most likely, not be accurate. For instance, if we would not have scaled our data, the model would have given the clusters of Figure 3.26 as the result of the mathematical analysis.

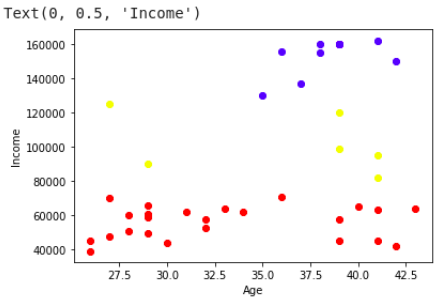


Figure 3.26: Clustering Error

We can see that, as we asked, the machine determined three clusters. However, there is something wrong with, at least, the yellow and the red cluster. In other words, the data points are not grouped correctly. In this case, the machine clustered the data points accordingly to the footballers' income but not to their age. For situations in which the income is a more important feature than the age, this result may be accurate.

3.3 The technical use of *learning*

How does the concept of learning work in machine learning? What does the computer scientist mean when applying the concept of learning to the models above? To which concepts is the concept of learning connected? Due to the complexity of the models, it is easier to answer these questions by having a panoramic view of the stages of each machine learning process. To begin with, in supervised learning, linear regression models, as the one described above, have the following phases.

1. Collection, ordering, and discrimination of relevant data.
2. Dividing the data into training and testing sets.
3. Specifying which is the label to be predicted.
4. Building the linear regression model.
5. Machine learning stage.
    (a) Training the model with the specified data. The model determines the patterns within the data, in this case, the best-fitting line.
    (b) Evaluating the model's learning performance and accuracy. The model is ordered to give the specified prediction for testing data.

In this case, the use of the concept of learning is related to the concepts of training, experience, evaluation and testing, induction, and prediction. The machine is *trained* when the expert gives correlated pairs of examples (training data), which function as general rules that are intended to guide the machine's performance. This is the *experience* the machine acquires (Mitchell 1997, p. 2). The machine *learns* when it determines a pattern or structure in the data, based on a particular mathematical model. By evaluating the machine, we say that it learned if it can predict the specified value or label. This prediction is also called induction. The machine gives a general statement or result based on having processed concrete data.

In the case of unsupervised learning the use of the concept of learning is slightly different. The machine learning process in clustering models, as described above, is the following.

1. Collection, arrangement, and scaling of the data.
2. Specifying the number of clusters the machine must find in the data.
3. Building the k-means model.
4. Machine learning stage.
    (a) The machine learns and determines the underlying structure and patterns of the data, by iterating the mathematical model until there is no variance in the clusters.
5. Plotting the results.

The use of the concept of learning in the case of clustering we described is, we may say, more circumscribed. By using this concept concerning the model above, the computer scientist refers neither to a training nor an evaluation process. The use of the concept of learning, in this case, is not related to the concept of experience. Rather, the use of *learning* is related to the concepts of finding, discovering, determining, etc. the structure of a certain data set automatically, that is, without the aid of training examples provided by the expert. In other words, the machine

learns when it assigns each data point to a specific cluster without human intervention. An important issue that needs to be addressed in this type of model is to what degree unsupervised learning is indeed unsupervised. To clarify this point we need to consider that the machine does not perform *autonomously*, in the sense that the machine does not perform without any kind of human intervention. The computer scientist designs the algorithm, orders the data, and uploads it while specifying in advance the number of clusters the machine must detect in the data set. Finally the expert implements the mathematical model so the machine can find patterns in the data set. Similarly to supervised models, the expert, we may say, is always behind the machine learning process.

In machine learning, as we can see, computer scientists make use of concepts like *learning*, *experience*, *reading*, *training*, *testing*, *inferring*, *finding*, and *discovering*, to describe how machines perform. We need to bear in mind that these concepts come from natural language. These concepts are far from having a straightforward meaning, and thus, describing their use does not represent a simple undertaking. Instead, these are heavily loaded concepts that convey a strong theoretical burden when imported to machine learning from their original field of use. The use of such concepts concerning machines needs to be handled with care and awareness. In this sense, describing the machine's performance with imported concepts from the human domain is problematic. For example, the familiar uses of *experience* do not point to the example data the machines use to make inferences. Rather, this concept is related to notions such as undergoing a given event, to the human sensual apparatus, to the knowledge someone acquires when living or going through a certain situation, or when practicing a specific activity for a long period of time, etc.

So far we have described some examples of human learning that are related to some of the features of the machine learning models described in this chapter. As the next chapters will show, the comparison between human learning and machine learning informs us in two directions, in the sense that, by clarifying how machines learn, we also gain valuable insight into how human beings learn.

Chapter 4

On the Similarities and Differences Between Human Learning and Machine Learning

In the last two chapters we described some examples of human learning and machine learning. Certainly, the use of the concept of learning is different depending on each particular context. However, human learning and machine learning maintain important affinities. What we want to say is that the use of *learning* in machine learning is analogous to the use of this concept in ordinary language. Two objects, words, situations, etc. are said to be analogous neither when they are identical nor when they are completely different. Rather, in terms of family resemblance, they are analogous insofar as they share significant features that allow their comparison. In this chapter we will show how human learning and machine learning are related by pointing to their similarities and differences. At the end of the chapter a table showing their similarities and differences is provided. This, in turn, will give us clarity to evaluate the application of the concept of learning to software and machines.

4.1 Similarities between human learning and machine learning

By having shown how both machine learning models work in chapter 3, we know that in each of these cases the concept of learning has a special use. That is, even in its technical application, the concept of learning is not univocal. In the examples described in the last chapter, computer scientists use this concept in two different, but related ways.

Let us remember the features of the concept of learning in the supervised model above. In this circumstance, the use of *learning* is related to concepts like training, experience, testing, and prediction. "Training the model" means that the expert gives the model the command to process pairs of correlated data points (x_1, y_1), that work as examples of correct associations. By using a mathematical model, the machine determines the underlying structure of the example data. In turn, by determining the structure of the training data, the machine is said to gain experience (Alpaydin 2014, p. 3). Thus, the machine acquires experience based on its training with example data. Finally, the machine is tested when the expert gives it the

command to predict or infer the value of a certain output (y_n) that corresponds to new, unseen input (x_n). If the prediction is satisfactory, the expert may say that the machine learned correctly.

This technical usage of *learning* is related, particularly, to the use of this concept in the familiar situation in which children are introduced to the primary colors. In this case, *learning* is related to the concepts of training, example, behavior, ability acquisition, testing, and teaching. Children are presented with samples of each color. The teacher asks them to utter the name of the colors several times. By doing this, children start to associate the word with the corresponding color. Once the teacher has enough evidence that children name the colors correctly, she tests their understanding by telling them to find, say, a blue object in the classroom. In this case, students do not learn by explanation. The teacher gives examples of correct courses of action. If children proceed correctly, the teacher may say that they did learn the colors.

The similarities between both cases are remarkable. What machine learning models can do is striking. In machine learning experts do not implement algorithms in the traditional sense. A machine learning model does not follow a sequence of instructions to produce a specific output given certain input. Likewise, children do not learn a procedure to determine colors: the teacher does not provide the wavelength and frequency for each color so children can determine which is the color of a certain object. In this sense, we can say that both machine learning models and children *learn* by examples. Machine learning models and children associate pairs of examples as right courses of action; "x_1, y_1" in the case of machines, and the color with the color-word in the case of children. With enough training in the case of children and example data or past experience in the case of machines (Alpaydin 2014, p. 3), both can associate correctly unseen instances to the correct color or label, respectively.

In contrast, the similarities between the use of *learning* in the unsupervised model above and in ordinary language decrease. The use of the concept of learning in our clustering model is not related to the concepts of training and experience (in the technical sense, i.e. *example data*). These models do not train, in the sense that they are not fed with pairs of labeled examples that show the correct associations and structure of the data. Our unsupervised algorithm does not make inferences based on previous experience or example data. Likewise, the use of *learning* in this model is not related to the concept of evaluation. Let us remember that the evaluation is made by the expert after the machine learning process finishes. In this sense, the concept of evaluation in our unsupervised model is independent of the concept of learning, which is applied directly to the machine's performance. In this case the use of *learning* is connected with concepts like exploring, determining and finding regularities, structure, patterns, etc. in the input data without the intervention of the expert. The machine is said to learn when it classifies or determines the location of each data point automatically in a certain cluster according to its attributes (Alpaydin 2014, p. 11, 173).

Unsupervised machine learning holds little or no relation to the situation in which children are introduced to colors. However, the technical use of *learning* in clustering is analogous to the use of this concept in the situation of the autodidact musician. In this less familiar case, the use of the concept of self-taught or autodidacticism conveys that the child has learned to play a certain instrument with little or no

formal instruction. Likewise, the concept of autodidact indicates that the child can identify the mistakes she makes in her learning process and correct them.

The similarities between the case of the autodidact and our unsupervised learning model are, first, that both the child and the machine are neither taught nor trained by a specialist. The machine does not count with previously labeled data that work as an example of correct associations. Likewise, the child does not follow the right course of action provided by the teacher. Rather, the child and the unsupervised machine learning model improve their performance without a direct intervention of the expert. The child can recognize whether she played the correct chord. Similarly, the unsupervised machine learning model improves the grouping of the data points into clusters automatically based on a mathematical model.

As we can see, there are significant criss-crossing similarities between human learning and machine learning. The usage of *learning* in machine learning is analogous to the usage of *learning* in ordinary situations. We may say, thus, that the way the machine learns is similar to the way children learn. In Hark's terms, "the behaviour of computers bears a certain resemblance to that of humans", in the sense that machine learning programs execute tasks that require intelligence in the case of children and adults (Hark 1990, p. 268). It is worth noting, however, that the previously mentioned similarities between machine learning and human learning do not depend on *what learning really is*. In other words, these cases do not point to the *true nature* of learning. As we have seen, there are countless instances of *learning*, which cannot be reduced to a single phenomenon or process (see section 2.1). Rather, the similarities make sense because of how the use of the concept of learning is structured. The grammar of *learning* as a psychological or mental concept allows the extension of its use beyond humans. This is worth considering. We will return to this point in the next chapter.

4.2 Differences between human learning and machine learning

Bearing in mind the uses of *learning* we have shown in the last two chapters, it is difficult to think that the concept of learning in machine learning can have (or could have) the same use as the concept of learning in natural language, that is, when it is normally applied to human beings. In other words, thinking that it is possible to apply the ordinary concept of learning to machine learning algorithms and software meaningfully is problematic. In short, machines and human beings do not learn in the same way. Human learning is a much broader instance of learning than machine learning. The concepts of learning, experience, training, inducing, generalizing, predicting, etc. in machine learning do not have the semantic weight they have when used to describe activities that human beings perform. In this sense, we could say that the use of *learning* in computer science represents a secondary use of the concept of learning in ordinary language (see PI, 282).[1]

The differences between both instances of learning are countless. We can always come up with a new example or situation in which the concept of learning is used

[1]See PPF xi, 276: "Here one might speak of a 'primary' and 'secondary' meaning of a word. Only someone for whom the word has the former meaning uses it in the latter." See also PPF xi, 278: "The secondary meaning is not a 'metaphorical' meaning."

to describe the behavior and actions of a human being to show the restrictedness of the use of *learning* in computer science. For this reason, it would not be fruitful to explore every difference we imagine. Instead, we will explore some of the most significant differences. It is worth noting that the following observations are closely interconnected. Their order is decided only for expository reasons, and does not represent any kind of hierarchy.

1. First of all, let us say that the use of the concept of learning concerning human beings *includes, or is connected with, a vast amount of elements*. Looking back at the different uses of this concept that Wittgenstein offers in the *Investigations*, we say that someone can learn: —a native language, —a second language, —the meaning of a word, —chess, —to apply a certain rule, —to lie, —algebra, etc. (See section 2.1)

Consider the case of learning algebra. What elements are connected in this situation? The child who is being introduced into this branch of mathematics needs to understand some elementary arithmetic operations, namely, she needs to know in advance how to count, to add, to subtract, to multiply, and to divide. This means that she needs to have certain knowledge about the use of some mathematical symbols, such as $+$, $-$, $\times$, $\div$, and $=$. She needs to be proficient in these practices before acquiring the ability to solve an algebraic problem. Another element that is taken for granted in learning algebra is that the child needs to be able to understand and follow the orders and rules the professor is giving. In other words, the child needs to respond in the ways we usually do when receiving a certain order. For this, the child needs to understand the language in which the professor is expressing herself.

Now let us consider the case of learning a new word by ostensive definition. Would we say that someone that has no knowledge whatsoever about what is a game —and therefore chess— understands the ostensive definition expressed like "*This* is 'the queen'"? It is natural to think that this question has a negative answer. Saying to someone, say, a newborn, that is not familiar with games at all "*This* is called 'the queen'" while pointing to the chess piece in question, would be senseless. For this expression to have meaning, that is, to achieve its purpose, it must be directed to someone who understands a language and that has a certain understanding of what the word "game" means. As Wittgenstein says: "an ostensive definition explains the use —the meaning— of a word if the role the word is supposed to play in the language is already clear" (PI, 30). Thus, the expression "*This* is 'the queen'" makes sense if the one that listens to it knows what a game is, what chess is, what a piece in a game is, etc. Likewise, expressions like "*This* is called 'maroon'", "*This* is called 'two'", and "*This* is 'jazz'", make sense if the interlocutor knows and understands the meaning of *color*, *number* and *music*, respectively. The point of these remarks is that human beings do not learn a certain skill *out of nowhere*. Certain knowledge, understanding, and skills are presupposed in —if not all— almost every learning process. In other words, we can say that in general, to understand a certain system —a certain language-game— one previously understands another one.[2]

[2]The chain of systems or language-games that are understood by someone that is learning a new skill, concept, or system ends with the complex case of learning a native language. This case does not presuppose certain knowledge, particular skills, or understanding, but primitive ways of reacting, a certain regularity in the child's responses. The discussion is related to the concepts of facts of nature and concept formation, among others. See Hertzberg 2011. See also Hacker 2012a, p. 208: "Initial language learning is training, which presupposes a wide range of common innate capacities, imitative propensities, and natural responses to stimuli."

In contrast, the machine does not *know* anything before performing a certain task, like playing chess or predicting the students' final grades. The machine's learning process is not related to any previous knowledge or skills. The use of *learning* in machine learning does not include a broad spectrum of elements. Instead, the use of this concept is related only to the outcome of the analysis of data using a mathematical model. In other words, before the learning process, the machine is built by an expert based on well-structured data and a corresponding mathematical model.

2. As we have seen in chapter 2, one of the ways to understand the meaning of learning is by saying —in Wittgenstein's terms— that this concept points to the different methods by which someone is "being brought to the point of being able to do [something]" (PI, 383). This means that the concept of learning encompasses the different ways in which someone is *introduced to a certain practice*. In other words, learning something includes acquiring the skills to play an active role in a certain activity or human institution.

For instance, learning how to count means preparing the terrain for being an active element within the activities and institutions in which this technique is applied. That is, learning how to count does not only mean acquiring the skill to reproduce the series of natural numbers *ad infinitum*. Instead, learning to count, one might say, opens the door for having an active role, for example, in using calendars, playing games, shopping, using money, establishing the number of people of a certain group, being able to keep track of one's age, generating a census, following the bar or keeping time in a musical composition, etc. Learning to count, therefore, enables the person to be a partaker, an actor in the practices in which this technique is used.

Acquiring the skill of counting means that the person can apply this technique in a wide range of practices. Now consider, one might say, a narrower case, namely, learning to ski. This technique is not as crucial as counting, which we use daily. Nevertheless skiing is a human institution. Hence, someone who acquires this ability is not just able to glide in snow, but to be a member of a cross-country team, to invest in equipment or to sell it, to take skiing as a leisure activity, to enjoy skiing, to admire a ski professional, to teach someone how to do it, to follow and apply the rules of a ski competition, etc. Thus, the person can enroll in this institution as an active element.[3]

In contrast, the concept of learning in computer science does not have the use above. Computer scientists do not use the expression "The machine learns" to indicate that the machine is acquiring a certain skill to be an active part of a human institution. In other words, the role the machine plays in our institutions is not that of an active member. In this sense, the machine is not an agent but an instrument we use to facilitate the achievement of a certain goal.

Consider the following example. IBM WATSON Beat is a machine learning neural network-based program that produces original music. To teach the system, computer engineers had to break "music down into its core elements, such as pitch, rhythm, chord progression and instrumentation." Later on, engineers "fed a huge number of data points into the neural network and linked them with information on both emotions and musical genres. [...] The idea was to give the system a set of structural reference points so that [the computer scientists] would be able to define

[3]This does not mean that the learner is always an active element of the practice in question. One can learn to play chess while not being involved in the institution.

the kind of music [they] wanted to hear in natural-language terms" (IBM 2017). On this basis, the system started to work on original compositions when it learned to link emotions with different musical styles and theory. Nobody questions the fact that this technology is extraordinary. Based on mathematical models and neural network structures, machines are indeed able to produce original music.

Notwithstanding, would we say that, when learning these data, the machine plays an active role in the institution of music? Think about whether the machine can enjoy a certain composition. In other words, does the system feel happiness or grief while listening to a certain genre? Is the machine able to learn how to appreciate a composition? Can the software teach someone how to compose music or play a certain instrument? Does the software have goals of its own to compose a piece of music? Based on the uses of these expressions, and the enormous amount of elements they include, we would not answer affirmatively to these questions. Once again, this software is not an agent but an instrument in the institution of music. The software helps and assists people that play an active role in the family of practices of music. Consider what people who work with this software say, namely, that the machine is "a creative assistant tool to help human composers rather than trying to replace them: you give the system a starting point, it comes up with a path to follow, and then you apply your own ingenuity and instinct to finish the piece" (IBM 2017).[4]

3. How do we use the concept of machine? The grammar of this concept demands much more than a few lines to be described. Suffice it to say here that a machine is an *instrument which is supposed to work*, and *to work satisfactorily*. The machine is efficient when it does what it is meant to do, or when it proceeds according to the way it was programmed. We assume the machine performs as it is supposed to. This means that we use the concept of machine evoking that, regarding its performance, this gadget or device has no choice. But not because this object does not have free will, but because it does not have the typical features, behavior, actions, and reactions of something or someone able to decide. In other words, if a certain machine does not respond according to the way it is programmed, we would not say "It decided not to work now", but simply "It does not work".

Now think about the concept of human being. In contrast with the remarks above, one of the multifarious ways in which we use this concept is to refer to someone that can decide whether or not to do something, that is, to act according to reasons. The grammar of both concepts, thus, is different. Machines and human beings behave in different ways. The proposition "x decides whether taking action or not" does not belong to the grammar of the concept of machine. Our reactions and behavior towards human beings and machines are qualitatively different (see PPF iv, 22; PI, 284).

In this sense, the machine has neither opinion nor choice about its learning process. Otherwise, this object would not be called "a machine". In other words, the machine does not act, it does not take an active part in the learning process. Instead, the role the device plays in the learning process is passive, that is, someone

[4]This observation is not beyond criticism. Someone might disagree about the degree of this difference by pointing that in some cases, machines do play an active role in human institutions. For instance, in the case of robots that assist the elderly in Japan (see Foster 2018). We would argue, however, that if agency versus instrumentality does not represent a radical difference between human learning and machine learning in this case, it is still problematic to say that the machine does not have an instrumental role whatsoever. At least, this observation remains open for further debate.

builds its structure and it receives the data to be processed. In the case of human learning, we sometimes say that the learner chooses, that is, the student can decide whether to continue with the learning process. Think, for instance, in the case of learning to ski mentioned above. Perhaps the learner decides to quit because she realizes that she does not enjoy this activity, or because she needs to focus on her studies.

We take a machine in general, and a machine learning model in particular, in Wittgenstein's terms, as "a symbol of its mode of operation" (PI, 193). This means that we normally take for granted that the machine will work as it is supposed to, that is, according to how it is programmed. This is how the concept of machine is regularly used. It is correct to say that we usually "forget the possibility of [the machine's] bending, breaking off, melting, and so on [...]; we don't think of that at all" (PI, 193). But it is not the case that we forget —and of course we also do not bear in mind— the possibility of its choosing or deciding whether it wants to perform a certain task. The language-games of making decisions, changing one's mind, etc. are not played in this circumstance. The learning the machine performs is circumscribed to the data it processes with the aid of a previously chosen mathematical model.

4. Consider the *machine's constitution*. How is the machine learning model composed? Which are the necessary elements for the program to run correctly? As we have seen in chapter 3, for building a basic machine learning program we need hardware, an operative system, a programming language, a mathematical model, and data. Some of these elements, by stipulation, have been distinguished as the external and internal components of the machine, i.e. hardware and software, respectively. With these elements and with the necessary knowledge to handle them, computer scientists use the concept of learning concerning the performance and workings of certain models. Regarding this set of elements and techniques, thus, the learning process of the model rests upon the machine's constitution. In other words, the use of learning in computer science encompasses a correspondence or connection among the constitution of the machine, its performance, and the learning process or outcome. If these traits or some of them would not be present, we would not have a learning process. The learning process, then, is the outcome of the presence and proper handling of these elements.

From a Wittgensteinian point of view, this way of using the concept of learning would be neither meaningful nor relevant if discussing the learning process of human beings. Although it is tempting to see the learning process of human beings somehow similar to the learning process of machine learning models[5], the use of the concept of learning in natural language does not include any kind of connection between so-called internal states or processes and the acquired ability in question. The use of the concept of learning in natural language is rather different.

[5]The picture some scientists and philosophers have, particularly in cognitive science, about the human learning process is that in some sense such process is similar to that of the learning process of machine learning-based models: the data comes from the outside and it is processed according to our internal cognitive structures (software, programming language, etc.). These structures lie somehow in our bodies (or hardware). We transform the data into experience, that is, we learn from the data provided by the environment. The similarities between cognitive theories of the mind and the actual working of machines deserve further consideration (see Walsh and Lovett 2016). This is also related to functionalism in cognitive psychology, in which the human mind is seen as a computer, an information processing system (Hark 1990, p. 267). Therefore, not all philosophers of mind would agree with this point. The discussion remains open.

First let us say that defending a correspondence between physiological (and/or mental) processes and mental concepts —such as thinking, imagining, understanding, learning— is highly problematic. Wittgenstein has many remarks that address these issues. Let us consider some of them.

> A misleading parallel: psychology treats of processes in the mental sphere, as does physics in the physical. Seeing, hearing, thinking, feeling, willing, are not the subject matter of psychology *in the same sense* as that in which the movements of bodies, the phenomena of electricity, and so forth are the subject matter of physics. (PI, 571)

> The psychological verbs to see, to believe, to think, to wish, do not signify phenomena. But psychology observes the phenomena *of* seeing, believing, thinking, wishing. (Z, 471)

Among other things, what these remarks are trying to show is that, in contrast to what happens with physical objects, we cannot point to a phenomenon to ostensively define a mental concept because *there is no such phenomenon*.[6] There is neither an object to point to nor to look for when investigating the meaning of these concepts. Thus from a Wittgensteinian point of view, no internal process, neither mental nor physiological, constitutes the meaning of the concept of learning. This is what Wittgenstein has in mind in the following remark.

> It is thus perfectly possible that certain psychological phenomena *cannot* be investigated physiologically, because physiologically nothing corresponds to them. (Z, 609)

In other words, no connection between, say, a brain state and the understanding (or the lack of it) someone displays when learning algebra seems to be necessary. More precisely, the meaning of learning as a psychological concept is not a *something* for which the word *learning* stands. Thus,

> No supposition seems to me more natural than that there is no process in the brain correlated with associating or with thinking; so that it would be impossible to read off thought-processes from brain-processes. (Z, 608)

Mental concepts are complex concepts. What we suggest by saying this is that the meaning of such concepts —as the meaning of many others— is the multifarious uses they have in our natural language (PI, 43). The concept of learning, thus, cannot be defined a priori, that is, we cannot give the necessary and sufficient conditions of its application because its use is not decided in advance. As we have seen in section 2.1, *learning* has different uses in different contexts. Such uses share some features but they may differ in other aspects. To understand the concept of learning we must look at the circumstances in which it is applied and describe its use.

[6] Thinking that the meaning of mental concepts are objects or phenomena leads to the problem of mental privacy, which entails important conceptual inconsistencies. This topic has promoted an extensive discussion (see PI, 243-315). We will not offer an account of this problem.

In the light of these remarks, let us contrast the case of error in the learning process of both machines and human beings. There are several types of errors in machine learning, for instance, using a single mathematical model to gain knowledge from different data sets, preparing the data inadequately, not defining a certain function, etc. For now, let us consider the first error in chapter 3. Once again, a model is said to be overfitted when it memorizes the patterns of the training data and is unable to predict the right output for new instances. In this case, the machine does not "learn" to generalize. This represents a problem because the goal of machine learning is not to repeat the structure of the training data, but to predict the result for new, unseen data. More technically, "we would like to be able to generate the right output for an input instance outside the training set, one for which the correct output is not given in the training set" (Alpaydin 2014, p. 38-39).

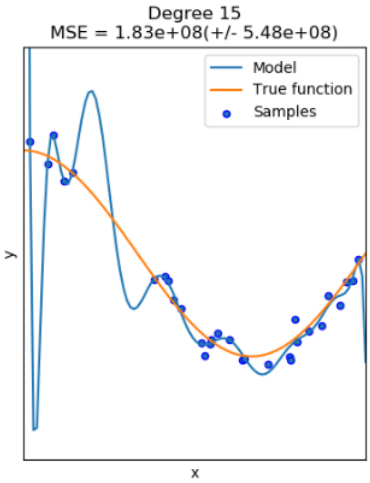


Figure 4.1: Overfitted Model

As we can see in Figure 4.1, by memorizing the exact location of the data points, the model determined the structure of the data set generating a complex function. Although the performance of the model with training data is almost perfect, its performance with new data will be unsatisfactory, because the true structure of the data set is a cosine function.

Overfitting has different causes. One may be that the mathematical model we chose is inappropriate, perhaps too complex, like in Figure 4.1. Another possible cause is that the training is performed with the entire data set. That is, when the data set is not divided into training and testing sets. Once we have determined the cause of the model's unexpected performance, we can prevent this malfunction by rearranging the machine's constitution, for instance, we can readjust the training and testing parameters, or modify the mathematical model.

Now consider the case of error in human learning. Imagine that a child is being introduced to chess. We show her the different pieces —"*This* is 'the queen', *these* are 'the bishops'"—, we show her how to place them in the board while saying something like, "*Here* goes the queen, *here* the knights and in *this* row you place all the pawns". After, we teach the child the rules of the game, for instance, the fact that a player can only move one piece at a time (except for castling), etc. Once the rules are clear for her, and once we have rehearsed a few times, the child starts

playing against us. Suddenly we realize that she made an illegal move by accident (see Figure 4.2).

Figure 4.2: Illegal Move in Chess

She played Kc2, that is, she placed the king into a square that is being attacked by a black piece, in this case the b3 pawn. We may say something like "No. Look closer. Remember that you cannot move the king to a square in which the king is in check."

In contrast with the case of the machine's malfunction, from a Wittgensteinian perspective we do not look for and point to the cause or causes of error in the child's case. We do not look into the child's head searching for the cause of her mistake. Would we say that the way for correcting her slip is to readjust her neural configuration? This is not how the concept of 'correcting a mistake' works in the case of human beings. An accidental illegal move in chess is not the direct effect of someone's brain state. In this sense, the child's behavior is not a matter of cause and effect. Rather, we would attribute this mistake to a lack of practice, attention or understanding, confusion, or perhaps to the fact that she forgot that particular rule. Thus, the ordinary ways in which we would correct the child's mistake are, for instance, (i) explaining to her why this is a mistake, (ii) using the rules of chess as a reason to pause the game, (iii) reminding her the general rules of the game, (iv) encouraging her to practice, (v) playing with her, etc. Similarly, when the child becomes aware that she blundered, and corrects her mistake, the expression "She learned the rule!" does not mean that she changed or readjusted her brain's configuration. Learning, as we have seen before, is understood as a change in behavior and ability acquisition. There are no internal structures that constitute the foundations of the use of the concept of learning (see Shanker 1998, p. 112ff).[7]

5. From a Wittgensteinian point of view and in close connection with the above, the *distinction between reasons and causes* plays a crucial role in understanding the differences between human learning and machine learning. Consider the following questions.

1. Why did the machine exhibit *this* result?

[7]This point remains open to debate. Some scientists and philosophers argue, for example, that the case of error in human learning is related to the cognitive architecture of the subject which can be represented by the structure that neural networks display. See Thagard 2012.

2. Why did you move the pawn like *this*?

These questions, on the surface, have the same grammatical structure. The use of the word "why" appears to be univocal in both cases (BB, 15). The form of both questions, according to Wittgenstein, promotes conceptual unclarity regarding the type of answer each question asks for. We are inclined to think that what these questions require is a *cause* as the explanation for the machine's and the child's performance and behavior, respectively. We may be tempted to reduce the justification for the child's actions to causes and causal explanations.

Think about the first question. First of all let us discard the fact that this question asks for a reason. The machine is not going to tell why it decided to display this result instead of that one. The machine learning model's performance is said to have *causes*, namely, a programming language, the data the machine processes, the mathematical model that allows the analysis of the data, etc. If the machine learning model's performance unexpectedly stops, we may have a certain hypothesis to explain this malfunction. Someone may say that the breakdown is due to a virus, or a glitch in the program, or maybe the data was not correctly processed. However, in the vast majority of cases in which we count with the proper knowledge and skills, it is entirely possible to find the cause or causes of such failure. Most likely, we can say that this inquiry aims to prevent future deficiencies. The grammar of "cause" is connected with the grammar of "hypothesis" and "explanation". In this circumstance, if we are unable to explain the malfunction we would not say that the machine's unexpected performance was due to an inexplicable cause. Rather, we would attribute our puzzlement to a lack of knowledge.

Now consider the case of human beings. As we have seen in (4), from a Wittgensteinian perspective it is deeply problematic to say that a certain physiological state, for example, a particular neural configuration, is the cause of our understanding a sentence, associating two concepts, seeing an aspect, learning to play chess, etc.[8] Imagine that in the child's learning process, we want to test her understanding of the rules of chess. We ask her "Why did you move this pawn two squares forward?" The why-type questions are sometimes misleading because they may intimate a hidden cause of human action. However, with this question we are not asking the child to give us the cause for her moving the piece in a certain way. The way she acts is not to be reduced to her physical constitution only. Suppose that the child does understand the rules of chess and replies something like "Because you taught me that when the pawn is moved for the first time it is possible to move it either one or two squares ahead." By expressing these words, the child has provided us with a motivation for her acting in this way. In her answer, a rule of chess is used as the reason she had to act as she did. She is not justifying her action based on what happens inside her head (or mind). As Wittgenstein says, "Giving a reason for something one did or said means showing a way which leads to this action" (BB, 14. See also Shanker 1998, p. 114, and Waismann 1968, p. 122).

It may be appealing to think about the learning process as building up a psychological mechanism, for example, a mental image of the rules of the game, which may be considered the cause of the child's actions. Nonetheless, this explanation

[8]This is not to say that we do not talk about causes regarding human beings. An example of this is the case of pain. Consider an athlete who is about to win the first place in the 100 meters Olympic competition. Suddenly she gets a leg cramp and loses. Here we do not use the concept of reasons. Instead we would say that the cause for her to lose was the cramp.

"would only be a hypothesis or else a metaphor", but certainly not the description of what the child does (BB, 12). This hypothesis may emerge from the conceptual unclarity explained above, namely, from confusing the grammar of reasons with the grammar of causes due to the form of the question. If we take this hypothesis as the cause of the child's action, then we would have an example of a cause that cannot be determined.

If we are aware of the differences between reasons and causes, on the one hand, and between both instances of learning on the other, then it is possible to make sound comparisons or analogies between the case of the child and the case of the machine. Paraphrasing Wittgenstein, it is possible to compare the child's learning process to the construction of a certain machine learning model. The parallel to the machine learning model's malfunctioning would be "what we call forgetting the explanation, or the meaning, of the word" (BB, 12).[9]

6. Mental concepts, as we have said, are complex concepts. Their use is not invariable. They present important *asymmetries*. An example of this comes up when we compare their use in the first person with their use in the third person of the present tense. Because the scope of mental concepts is immensely wide, concepts of sensation, perception, and emotion are not going to be considered. Instead, we will analyze two instances of what we would like to call —in contrast with the former classification— intellectual concepts (see Hark 1990, p. 198). Consider the following expressions.

1. *Fania knows* how to code in Python.
2. *I know* how to code in Python.

How does "knowing" work here? Let us describe the use of this concept in the expressions above. From a Wittgensteinian perspective, the meaning of our mental concepts is their use in language and not a phenomenon for which our words stand. The use of "knowing" in the third person of the present tense —"Fania knows how to code in Python"— does not mean making an inference from the observable behavior to the mental, internal states of a certain individual. Instead, this use provides information about someone based on certain evidence (RPP II, 63). For example, Fania may have told us that she took a course on that particular programming language. Or perhaps we say this because we have seen her using Python. Maybe she is a teacher or a professional software developer. On the other hand, "knowing" in the first person of the present tense —"I know how to code in Python"— is neither used to describe nor report what happens inside our minds. This use of "knowing" does not point to an internal state that only I can have access to. This expression "does not rest on introspection conceived as inner sense; nor does it rest on observation of one's own behaviour" (Hacker 2012a, p. 209). In this sense, this expression does not give information about our mental state. The use of this concept in the first person —in contrast to the use in the third person— does not rest upon any kind of evidence. We do not look into our heads to be sure whether we know something. Rather, this use of the concept of knowing is similar

[9]This topic is not exempt from philosophical disagreement. As in (4), the cognitivist may argue that human action is based on causes. Davidson 1963 holds a critical posture towards the points of view expressed here by holding that reasons are the causes for action. For a comparison between Davidson and Wittgenstein about the reasons and causes debate see Glock 2014.

to an expression (RPP II, 63). But an expression of what? Consider the following remark.

> The grammar of the word "know" is evidently closely related to the grammar of the words "can", "is able to". But also closely related to that of the word "understand". (To have "mastered" a technique.) (PI, 150)

That is, the use of "knowing" in the first person of the present tense refers to the announcement or declaration that we can do something. Imagine that a computer science student needs help with her programming assignment and Fania replies "I know how to code in Python!" What she means is that she understands the programming language in question, she is expressing that she can offer help to the student, etc. The differences between both uses, then, need to be regarded. Human learning constitutes an instance of this asymmetry. Let us offer an example.

1. *Fania is learning* to code in Python.
2. *I am learning* to code in Python.

Similarly to the case above, the use of *learning* in the third person of the present tense does not depend on an inference from Fania's behavior to her internal, mental states. The use of these words depends on certain evidence, for instance, that Fania takes a certain programming course, that she studies computer science, that we see her while programming in Python, etc. The use of *learning* in the third person of the present tense provides information about the person, for instance, that her behavior regarding the use of computers is changing, that she is becoming able to code, that she can build more complicated models, that she is gradually acquiring a better understanding of the system, that she gradually incorporates to her vocabulary a new terminology, that she is starting to develop programming skills, that she can understand and explain complicated topics in computer science, etc.

In addition to this, the use of *learning* in the first person of the present tense is not a report of an inner, mental state. This use of words offers neither information nor facts about our cognitive apparatus. The sentence "I am learning to code in Python" may be an expression of satisfaction when building a model that works for the first time after several unsuccessful attempts. In this sense, this expression is an avowal and does not depend on evidence. We realize we are learning, one would like to say, immediately.[10]

Let us now ask how the concept of learning is applied to software. The complex structure displayed by the use of these mental concepts in natural language (i.e., when applied to human beings) does not apply to the technical use of the concept of learning. First and foremost, it is evident that the machine learning model does

[10]However, there are utterances in the first person which can be considered as descriptions of one's own state. But Wittgenstein regards these expressions as a rather small group of the first-person statements. "Describing my state of mind (of fear, say) is something I do in a particular context" (PPF ix, 79. See also RPP I, 693). This particular context may be, for instance, "the psychological laboratory, Freud's sofa, Proust's bed" (Hark 1990, p. 115). For a deeper treatment of the classification of our psychological concepts see Schulte 1993, chapter 3.

not make use of mental concepts in the first person of the present tense.[11] Because of this, the analysis of the concept of learning in the first person regarding the machine learning models we described is discarded. Nevertheless, we indeed talk about the machine's workings and performance. Regarding machines, thus, the use of "learning", one would like to say, makes sense only in the third person. Consider the following use of words.

– *The machine learns* to classify the data.

This expression gives information about the machine's performance. This use of words refers to the current workings of the computer: at this stage, the machine is working with the data. This expression is based upon (i) the knowledge of the expert about the mechanisms and functioning of the machine learning model and (ii) the behavior the machine displays in certain circumstances. The expert emits this utterance by observing the model's performance (RPP II, 63).

In this sense, there seems to be a strong parallel between the technical and the ordinary use of learning in the third person of the present tense. Both uses are based on the observation of the machine's and the amateur programmer's behavior, respectively. However, this parallel is valid just in a restricted sense. Expressions like "The machine is learning to classify the data" make sense in the context of data analysis and machine learning. That is, this expression finds a use because the expert knows what happens in the machine learning model. The ordinary use of learning, in contrast, is not restricted to a specific domain: every competent user of natural language knows how to apply this concept in an open-ended number of contexts.

The parallel also seems to vanish when we consider the elements that contribute to the usage of each expression. As we said, "Fania learns to code in Python" is not an inference we make from the individual's behavior to her mental states. In contrast, the expression "The machine learns to classify the data" seems to depend on an inference. Based on the result or output the computer displays, the expert infers whether the machine is applying correctly the mathematical model to the data set. Based on this evaluation, the expert may decide to analyze another data set with the same machine, or readjust certain parameters to make the machine more efficient.[12]

[11] This point is parallel to the discussion about the ascription of mental concepts to some animals. When exploring the ascription of thinking to apes, Wittgenstein points out that the use of this concept "doesn't have a first person present" (RPP II, 231), and thus "Not until it finds its particular use in the first person does it acquire the meaning of mental activity" (RPP II, 230). According to Shanker, "The absence of the first-person use of the verb 'to think' —which is an integral aspect of the concept of thought— underscores the limited sense in which the chimpanzee can be said to think. Thus we might refer to the chimpanzee's behavior as exhibiting a 'primitive' use of the verb 'to think'" (Shanker 1998, p. 158).

[12] Just as some of the observations above, these points can receive critical assessment. The perspective according to which some mental concepts —such as knowing— represent *propositional attitudes* —which can be defined as the "relations between organisms and internal representations" (Fodor 1978, p. 501)— is an example of a philosophical view that does not consider the asymmetry that some mental concepts display in their ordinary use. According to this view, a propositional attitude is a cognitive relation that an individual has with a certain proposition (see Fodor 1978, p. 502). In expressions like "She knows P" and "I know P", the verb "to know" means the same, i.e. there is no crucial difference in its use that needs to be regarded. This view, then, presupposes a generalization in the use of this mental concept, which is problematic. See also Churchland 1981.

4.3 The different uses of *learning*: a comparative table

To gain clarity of the logical landscape of learning, Table 4.1 displays the comparison among the different examples of learning that we have described in past sections. This table exhibits the relationship between each instance of learning and some of the most relevant concepts we considered. The visual representation of the similarities and differences among the uses of learning we described will facilitate the understanding of this concept. Number *1* means that the type of learning is normally related to the concept in question, *0* means that this type of learning is normally not related to such concept, *1/2* means that we think of the type of learning as related or not related to such concept, and x means that the type of learning is independent of that concept.

Table 4.1 helps us to understand that *learning* has no essence, that is, that this concept cannot be reduced to a general definition that gives an account of every one of its different instances. Instead, this table shows that the different examples of learning share important criss-crossing similarities in virtue of which we use this concept (PI, 65. Also see Glock 1996, p. 121 and Forster 2010, p. 69). *Learning* is a family resemblance concept which has an open-ended number of uses. The table shows the differences in the meaning between the technical and the ordinary concept of learning. In this sense, we can say that different but related language-games are played with these concepts (see Shanker 1998, p. 199). Regardless of the differences between human learning and machine learning analyzed in the last section, this concept can indeed be applied to machines meaningfully. The different uses of "learning" in machine learning described in chapter 3 represent a continuation of the family resemblance aspect of this concept. Concerning the comparison between human and machine learning, Table 4.1 helps also to understand that the similarities between these learning instances would seem more striking if we overlook the role the human has when operating the machine. We must be aware of how much the human is involved in the practice of machine learning.

The use of the concept of learning concerning machines is legitimate. As we will show in the next chapter, however, what is not legitimate is to ignore the differences among the manifold instances of learning, for example, if we interpret the use of this concept concerning machines under a mentalist perspective, or if we reject the use of this concept concerning machines because machines do not have what we call a mind.

Related concepts	Learning the colors	Self-learning	Supervised learning	Unsupervised learning
Experience	0	1/2	1	x
Training	1	1	1	x
Evaluation	1	1/2	1	x
Induction	0	x	1	x
Change in behavior	1	1	x	x
Ability acquisition	1	1	x	x
Examples	1	1	1	x
Authority figure	1	1/2	1	1/2
Memory	1/2	1	1	1
Doing the same	1	1	x	1
Understanding	1	1	x	x
No formal instruction	0	1	x	x
Imitating the behavior	1	1	x	x
Auto-correction	1/2	1	x	1
Deciding on the learning process	0	1	x	x
Willpower, tenacity	0	1	x	x
Mathematical model	x	x	1	1
Data	x	x	1	1
Detecting patterns automatically	x	x	x	1

Table 4.1: Comparative table of the uses of *learning*

Chapter 5

On the Sense of the Application of Mental Concepts to Software

Within the field of AI, as we have seen, computer scientists make use of mental concepts and expressions to evaluate and talk about the computer's performances and functioning. This is still a matter of philosophical debate. Numerous texts are concerned, for instance, with the problem of whether a machine can perceive, think, or understand. In the past chapters we have explored some examples of human learning and machine learning, as well as the relations that hold among them. This analysis can clarify under which circumstances machines can be bearers of mental concepts. In other words, the analysis of our past chapters can shed light on philosophical problems concerning the application of mental concepts to machines and software. This is the general objective of this chapter. In the following, we analyze three different conceptions about the use of mental concepts concerning machines and software. The first one is a variation of what has been called strong artificial intelligence (see Searle 1980 and Chalmers 1996). The next section contains the analysis of what we call the rigid conception of grammar, the view that rejects the use of mental concepts concerning machines. The third section is dedicated to the perspective we defend, namely, that the sense of the ascriptions of mental concepts to software is interconnected to the practices they are embedded in. It is worth noting that the use of mental concepts concerning machines is something given. The following analysis starts from this fact.

5.1 A philosophical usage of mental concepts

Consider the following quotations.

> We can tentatively define a superintelligence as *any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.* (Bostrom 2014, p. 22)

> one day (as we have suggested) [machines] will be superintelligent. (Bostrom 2014, p. 22)

> True superintelligence [...] might plausibly first be attained via the AI path (Bostrom 2014, p. 50)

> A machine superintelligence might itself be an extremely powerful agent, one that could successfully assert itself against the project that brought it into existence as well as against the rest of the world. (Bostrom 2014, p. 95)

> it may be reasonable to believe that human-level machine intelligence has a fairly sizeable chance of being developed by mid-century, and [...] that it might perhaps fairly soon thereafter result in superintelligence; and that a wide range of outcomes may have a significant chance of occurring, including extremely good outcomes and outcomes that are as bad as human extinction. (Bostrom 2014, p. 21)

> If some day we build machine brains that surpass human brains in general intelligence, then this new superintelligence could become very powerful. And [...] the fate of our species would depend on the actions of the machine superintelligence. [...] Once unfriendly intelligence exists, it would prevent us from replacing it or changing its preferences. Our fate would be sealed. (Bostrom 2014, p. vii)

Several things can be said about this from a Wittgensteinian perspective in general. Let us just focus on two, interconnected aspects of the quotes above that seem to be the most relevant for our present purposes. (i) As the reader can notice, *Bostrom's claims consist of ascriptions of mental concepts to machines and software*. Intelligence, cognition, agency, intentionality, the capacity of adapting to different environments (learning), the capacity of establishing friendship or a hostile relationship with humans, are some of the concepts the author uses to describe the features of his hypothetical machine.

It is important to note that this author does not take into consideration that a possible way to proceed in the analysis of the philosophical problem of thinking machines is to regard the use of mental concepts as a signpost —immediate clear indication— of their meaning. In consequence, the author fails to notice that the use of these concepts is multifaceted. As we have seen in the case of the concept of learning, the meaning of mental concepts is not obvious and, due to its complexity, their grammar can be misleading.[1]

In this sense, the author assumes that the mental concepts he uses can be applied to machines in the same sense as these concepts are applied to human beings in normal circumstances. In other words, Bostrom employs these concepts neither with nuances nor restrictions, assuming that the meaning of mental concepts concerning machines is the same as the meaning of mental concepts concerning humans. The expression "A machine is (might be) an agent" is taken literally, i.e. it is taken in the same sense as the expression "Fania is an agent". Both expressions are considered equivalent, in the sense that their grammatical subjects ("machine" and "Fania") can be substituted for one another and the meaning of these expressions remains the same.

[1]For example, grammar can mislead someone who takes the form of the expressions as communicating something (metaphysically) true about the concepts used in them (see Z, 458). "I have a guitar in my room" and "I have a pain in my knee" display, at the surface, the same grammatical structure. One of the differences between both expressions, however, is that in the second one the verb 'to have' does not convey a relationship of possession. For an example of a misleading use of the concept of imagination see PI, 518.

Bearing in mind the last chapter's discussion about the differences between human learning and machine learning, we can say that Bostrom ignores the differences between the use of mental concepts concerning machines and the use of mental concepts concerning humans. This makes him overlook the grammatical differences between the concept of human being and the concept of machine, the context and background of the use of mental concepts when applied to humans, for instance, skills and primitive reactions, the role that training plays in concept acquisition, the human institutions and practices that surround psychological expressions, etc.

It is important to emphasize, however, that this overlooking does not take place in practice, i.e. in the situation in which the computer scientist makes use of mental concepts when building and operating the machine learning algorithm. When expressing "The model *learned* the best-fitting line for the training data" the expert does not mean the same as when she says "Today my daughter learned the colors at school". The use of *learning* in the first expression is circumscribed to a particular context and it serves a certain purpose, for example, reporting the performance of the machine, communicating that certain patterns have been localized in the data set, etc. From a Wittgensteinian perspective, we may say that the technical expressions of the expert have indeed a use, they are embedded in a particular scientific practice, while the expressions about hypothetical intelligent machines, do not. The latter expressions do not play a role in any scientific practice, and this means that these expressions do "not exist among our applications of language" (RPP II, 201).

In connection with this last point, the second aspect to consider about the quotes above is (ii) *Bostrom's point of view* on superintelligent machines, namely, that they will be "created relatively soon after human-level machine intelligence", which is defined as "one that can carry out most human professions at least as well as a typical human" (Bostrom 2014, p. 20, 19). Bostrom also maintains a "polarized outlook on the consequences [of creating a superintelligent machine], thinking an extremely good or an extremely bad outcome [such as human extinction] to be somewhat more likely than a more balanced outcome" (Bostrom 2014, p. 20).

Bostrom's perspective about these matters is supported by "a series of recent surveys [that] polled members of several relevant expert communities on the question of when they expect 'human level machine intelligence' to be developed" (Bostrom 2014, p. 19).[2] In other words, the author's view is based on the opinions of AI experts (Bostrom 2014, p. 18ff). *Prima facie*, one may think, experts know better, and thus, their opinion about these matters must be reasonable. What experts have to say about a hypothetical superintelligent machine, we may say, makes more sense than what we neophytes could say about the same topic. After all, we do not belong to the expert community, and this means that we lack important knowledge in the relevant field, in this case, AI.

Nevertheless, it is important to note that in this case we are not dealing with a scientific problem properly speaking, i. e. we are not dealing with a question that requires knowledge to be solved. Scientific problems in AI are, for example, detecting specific patterns of a certain disease in the human body using a machine learning algorithm, or translating a certain text based on previously processed data.

[2]It is important to mention, although this point will not be further elaborated, that the opinions of these experts, in turn, rest on strong assumptions. One of the surveys inquired the experts "about how much longer they thought it would take to reach superintelligence *assuming* human-level machine is first achieved" (Bostrom 2014, p. 20. My Emphasis).

The question of whether machines will be intelligent and unfriendly agents, instead, constitutes a conceptual problem. These problems can be solved by anyone who is a competent user of natural language —which does not mean, however, that these problems can be solved easily or that are not important. To clarify this issue let us consider a parallel case.

Imagine that a mathematician, an expert in arithmetic holds the view that "numbers are abstract entities, which exist independently of the human mind". According to Wittgenstein, "what [this] mathematician is inclined to say about the objectivity and reality of mathematical facts is not a philosophy of mathematics, but something for philosophical *treatment*" (PI, 254). In other words, what the mathematician states about the nature of numbers must be differentiated from the actual mathematician's work in her field, in this case, arithmetic. The idea that "numbers are abstract entities" represents neither an extension in our mathematical knowledge nor a proposition that can be proved or refuted by any mathematical method.

Similarly, we must distinguish between knowledge or advancements in AI and statements about a machine's hypothetical mental powers. "Machines will be superintelligent agents" is a proposition that does not represent scientific progress in the field of AI. Therefore, stating or not stating this does not affect the scientific progress in this field. In the same sense, "The superintelligent machine will be unfriendly" is a proposition that has the same epistemic value as, for example, "The superintelligent machine will be indifferent towards humans". These propositions are not based on empirical research, and no method or theory can prove its truth or falsity. "Machines will be superintelligent agents" only represents what the AI expert is inclined to say (see, PI 386). In turn, having and inclination or predisposition does not mean "having an immediate insight into, or knowledge of, a state of affairs", and thus, we are not compelled to accept the AI expert's point of view (PI, 299).

What the mathematician says about the nature of numbers and what the AI expert says about a hypothetical superintelligent machine are not scientific propositions. The truth or falsity of the latter can be proved or tested, they are based on empirical evidence, a certain (scientific) theory can give an account of them, etc. Both types of propositions do not stand together, and their difference is a matter of kind, not a matter of degree. If we do not belong to the expert community in mathematics or AI, then, the work of the experts lies beyond our judgment. The expert community sets the criteria for and evaluates the progress in each field. Arguably, then, philosophical work is circumscribed by what science can prove and explain based on evidence.[3] In contrast, we do can point to the inconsistencies and conceptual problems in the expert's ideas about her work. To explain this point it is useful to consider the following remarks.

> I am proposing to talk about the foundations of mathematics. An important problem arises from the subject itself: How can I —or anyone who is not a mathematician— talk about this? What right has a philoso-

[3]Philosophy and science are not continuous, but this does not mean that they conflict. The aim of Wittgensteinian philosophy is not to interfere with real scientific progress. Rather, one of its objectives is to scrutinize and judge statements of scientism. Important matters about these issues remain to be considered. For instance, the social, political and economic impact of claims stated in scientific terms by public figures without empirical evidence, the degree to which these conceptual assumptions can obscure scientific research, etc. See Tejedor 2017.

> pher to talk about mathematics? [...] I can as a philosopher talk about mathematics because I will only deal with puzzles which arise from the words of our ordinary everyday language, such as "proof", "number", "series", "order", etc. (LFM 13, 14)

> It is not the business of philosophy to resolve a contradiction by means of a mathematical or logico-mathematical discovery, but to render surveyable the state of mathematics that troubles us (PI, 125)

In other words, we are entitled to emit judgments about the use the expert gives to her words, because the language she uses is the same as ours. We both are users of natural, ordinary language. The expert, when giving explanations, she uses the language we use; her "questions [are] framed in this language; they [have] to be expressed in this language" (PI, 120).[4] In this sense, and based on the analyses in chapter 2, 3, and 4, we can say that the variety and richness of the vocabulary of the mental should warn us not to take Bostrom's claims uncritically. According to our methodology, at least some of his claims about these matters are beyond the bounds of sense.

The method of analysis of problems concerning strong artificial intelligence is not obvious. These problems are connected with questions and puzzles about the human mind, a topic which displays enormous dimensions and an inherent intricacy. Someone who approaches the discussion of strong artificial intelligence without having considered that the meaning of mental concepts is their use in language might be misled by important aspects of the use of our mental concepts.

5.2 A rigid view of grammar

Wittgenstein writes:

> a rule of grammar [...] *determines* the position of the word in language. (BT 199. My emphasis)

> Grammar *determines* the meaning of words (BT 198v. My emphasis)

Putting aside both the fact that Wittgenstein decided to make extensive corrections to what has been called *The Big Typescript*, and the fact that we do not find these type of expressions in the *Philosophical Investigations*, it is possible to take these remarks as hints for interpreting what Wittgenstein understood by "grammar" and how grammar establishes the meaning of words. We call the perspective according to which the meaning of words is determined by grammar, *the rigid view of grammar*.[5] Consider the following quotes.

[4] Advancing an investigation on the use of concepts in ordinary language is not endorsing what most of the people say. This would be a misconception of this type of philosophy. The idea of thinking machines, while being an actual possibility for a large number of people these days, rests on a conceptual incomprehension, just as the expression "Only I can know what I am thinking", which in general is accepted to be true in the common imaginarium.

[5] Arguably, it seems that in the *Investigations* Wittgenstein takes distance from this type of expressions. Most of the time he uses the verb "to determine" to picture philosophical postures he is criticizing. "Is my use of the name 'Moses' fixed and determined for all possible cases?" (PI, 79). See also PI 99, 136, 189, 193.

> The meaning of a word is determined by its grammar, i.e. by the familiar, accepted, rules for its use. [...] The location of a word in grammar is its meaning —its position in grammatical space. (Hacker 2012b, p. 6)

> Grammar [...] consists of sense-determining rules of a language. What belongs to grammar in this sense is everything required for determination of meaning. (Hacker 2012b, p. 5)

> Grammar [...] consists of rules for the use of signs that determine their meaning. (Hacker 2012b, p. 9)

> The a priori nature of things is fixed by the sense-determining rules for the use of expressions signifying things.[6] (Hacker 2009, p. 138)

> The grammar of language is a normative structure and speaking a language a normative practice. (Hacker 2012a, p. 220)

These statements suggest that the uses of concepts that are outside grammatical space lack meaning. In other words, if a certain concept is not applied following its normal, familiar rules of use in natural language, then this concept is not being applied meaningfully. The new use constitutes a deviation from the ordinary applications of the concept. It is in this sense that grammar determines which uses make sense and which uses do not. If we pay enough attention to the grammatical space of a certain word, we will know whether this or that application makes sense. In this respect, grammar, by "[laying] down rules that determine what makes sense", can be considered "a normative description (and investigation) of language" (Hacker 2012b, p. 4-5).[7]

The rigid view of grammar has important implications in our general understanding of the use of mental concepts and expressions. Hacker maintains that the application of mental concepts does not constitute an exception, but takes place under this conception of grammar. Concepts such as "know", "believe", "thought", "understanding", etc. "are constituted by the sense-determining rules for the use of the words that express them. [These rules] determine the meanings of the words we use"[8] (Hacker 2009, p. 143). Consider the following remark.

> It *makes sense* to ascribe a psychological attribute to another being, truly *or* falsely, only if it is *possible* for that being to display such behaviour as *would* count as good evidence for the ascription of the psychological attribute, i.e. the appropriate forms of behaviour must be in the creature's behavioural repertoire. (Hacker 2012a, p. 210)

[6] In the corresponding footnote, Hacker distinguishes between *essence* and *nature*, stating that family resemblance concepts, in this sense, do not have an essence, but may have a certain nature. However, the questions "What is the nature of games?" and "What is the essence of games?" do not seem to be very different. Games, we would maintain, do not have a nature.

[7] It is interesting to see the parallels Hacker draws between the grammar of language and jurisprudence and law. For example: "Philosophy is a tribunal of sense" (Hacker 2012a, p. 221).

[8] Hacker continues by saying that "It is important not to conceive of such rules in too formal a manner —we are not dealing with the rules of a calculus, nor yet with regimented grammar or lexicography" (Hacker 2009, p. 143). Passages like this do not seem to suggest that grammar is rigid. Nevertheless, Hacker still uses the concept of "to determine" to describe the features of the rules of language.

Based on his use of the conditional, Hacker's formulation can be rephrased saying that *if a creature cannot display the appropriate forms of behavior, then it makes no sense to ascribe a psychological attribute to it.* The meaning of mental concepts is determined based on, or according to the behavior of the creature in question. The familiar, meaningful expressions in which mental concepts are used are expressions related to human beings. Hacker maintains that the rules for the use of mental concepts are closely connected with the behavior of human beings. The ascriptions of mental concepts, then, depend on the behavior displayed either by a human or by a creature whose behavior resembles that of a human.

According to this view, ascriptions of mental concepts to human beings and some animals are inside grammatical space, i.e. they constitute meaningful applications, because the behavior of these creatures allows such ascriptions. Dogs, for example, count with the behavioral repertoire which allows us to ascribe some mental concepts to them. Their behavior is similar to the behavior of human beings in some circumstances. We may say that a dog is happy, or feels regret for having done something forbidden, etc. In contrast, ascriptions of mental concepts beyond human beings and creatures whose behavior is similar to that of human beings, are said to be outside grammatical space, and thus they lack meaning. Animals that display a significantly different behavior from ours, cannot be said to be bearers of psychological predicates. Expressions such as "fishes think" "[do] not exist among our applications of language", and this means that these expressions do not have a use (Z, 117). Let us consider the following remark:

> only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious. (PI, 281)

Hacker seems to interpret the passage above broadly, that is, that psychological attributes in general —and not just sensation and consciousness concepts— can be ascribed only to humans and to creatures that behave like humans. Thinking, understanding, following rules, knowing, etc. can be ascribed to creatures insofar as they count with the appropriate forms of behavior. On this basis, the question "Could a machine think?" (PI, 359) and a positive answer to it would lie outside the bounds of sense. We can almost reject this question and a positive answer to it a priori. The machine learning algorithms in chapter 3 do not display the typical behavior of a human being. Assuming that the rules for the use of mental concepts state that they can be ascribed only to human beings and to creatures whose behavior is similar to human behavior, then, these machine learning programs are excluded from such ascriptions.

Nevertheless, these matters are not that obvious. Mental concepts do not constitute a uniform mass, which functioning can be explicated by giving a straightforward account. *Pace* Hacker, we do ascribe mental concepts beyond humans and creatures whose behavior resembles human behavior, without ascribing, in turn, a mind in these circumstances. Let me explain this point with some examples.

We may ascribe, for instance, calmness or joy to drawings such as Figure 5.1. "In some respects, [we] engage with it as with a human face" (PPF xi, 119). These pictures can be helpful to teach children concepts such as happiness. We may also ascribe thinking to it, and children may also talk to it as they talk to dolls (PI, 360). We ascribe emotions to artworks, for instance, wistfulness and sorrow to Mary

in Michelangelo's *Pietà*, or thinking and deliberation to Rodin's *Le Penseur*, not because we ascribe a mind to the sculptures, but because the sculptures represent a tragic scene, a dejected mien, or express immersion in deep thought, etc. Moreover, these objects cannot be said to display any kind of behavior, neither similar nor different to the typical behavior of a human being.

Figure 5.1: Picture-face

Notwithstanding, we understand these expressions because they constitute familiar uses of mental concepts. These expressions and ascriptions are not outside grammatical space, and thus they are not excluded from grammar. On the contrary, the grammar of mental concepts allows us to use them in relation to humans, animals, drawings, paintings, sculptures, dolls, etc. Mental concepts are pliable. They have an open-ended number of uses. As Wittgenstein states: "these extremely general terms [mental concepts] have an extremely blurred meaning. They relate in practice to innumerable special cases, but that does not make them any *solider*; no, it makes them more fluid" (RPP I, 648).

Something similar applies to the case of the use of *learning* in computer science and machine learning. "The model learned the patterns" is a perfectly meaningful, though technical, expression. While operating, experts ascribe learning abilities to machines, neither because they ascribe a mind to them nor because the behavior the program displays is similar to human behavior, but because machines are built to perform in certain ways and produce a certain kind of results. As we have seen before, in supervised learning for example, the expert *trains* the model by giving it the command to process pairs of correlated data points, which work as instances of correct associations. Based on this *experience*, the expert tests the machine by giving it the command to *infer* the value of a certain output which corresponds to input the machine has not processed before (see section 4.1). Ascribing learning abilities to the machine means that the machine's performance is similar —in a restricted sense— to the behavior of humans when learning certain abilities.

Hacker's conception of grammar as determining the meaning of mental concepts based on human behavior excludes ascriptions of mental concepts to some objects that do not behave like humans (or that do not display behavior in general). Certainly and following Hacker, human behavior can be considered as a reliable standpoint and a powerful criterion for the ascription of mental concepts to some creatures, i.e. to creatures whose behavior resembles human behavior. The problem is to say that the *only* criterion for these ascriptions is human behavior.[9] To state

[9]According to this point, Hacker's view can be considered as giving a successful account of ascriptions of mental concepts to some machines and robots whose behavior is similar to human behavior.

this a priori is problematic. The criteria may well be open-ended or may change in time. In this sense, the rigid view of grammar may prevent us from seeing actual familiar and/or technical applications of concepts, or may lead us to discard some expressions as spurious. This view of grammar restricts language use on a particular basis (human behavior in the case of mental concepts) and fails to give an account of actual uses that conflict with such a basis. In this sense, this conception of grammar is not descriptive, but normative. Philosophy, however, "must not interfere in any way with the actual use of language" (PI, 124).

Furthermore, to "compare the use of words with games [is not to] say that someone who is using language must be playing such a game" (PI, 81). Similarly, to clarify the meaning of words by reference to sense-determining rules of use, does not imply that meaning can be explained only by reference to such rules. In other words, clarifying language use based on sense-determining rules does not imply that language, in general, is rule-determined. We must distinguish our mode of representation from the phenomenon we want to represent (PI, 50, 104). Kuusela puts this point saying that "it does not follow from the possibility of describing language in terms of rules —which Wittgenstein of course accepts— that such rules constitute the foundation of language" (Kuusela 2019, p. 147n).

No formulation of language and meaning can give account of all of their forms. This is not to say that certain formulations of language and meaning cannot give an account of certain expressions and uses.[10] Thus, the idea of grammar and meaning cannot be reduced to be the collection of "sense-determining rules of a language". The rigid conception of grammar, nevertheless, is useful to explain certain uses of concepts. In mathematics, for example, we can say that the grammar, i.e. the rules of use of signs such as $=$, $+$, $\times$, etc. determine what they mean. Based on the inflexibility in the use of these terms, we could say that they have a determinate meaning (RFM I, 1). But if we extend the rigid view of grammar to give an account of human language in general, this perspective will be prescriptive, and thus inconsistent according to some of our actual uses of language.[11] The grammar of mental concepts is, taking Kuusela's expression, multidimensional. The meaning of these concepts is not decided in advance, and cannot be explained before having attended to their actual uses.[12] I think it is in this sense that Wittgenstein suggests that "[o]ne cannot guess how a word functions. One has to look at its application and learn from that" (PI, 340).

To recapitulate. On the one hand, Bostrom's ascriptions of mental concepts to superintelligent machines presuppose the ascription of a mind to such machines. His perspective does not distinguish between ordinary and technical uses of mental concepts concerning machines, and thus his view about a hypothetical superintelligent

[10]This is the case of the Augustinian conception of human language (PI, 1ff). This conception does give an account of some instances and uses of our language. But if this conception is said to give an account of all the forms of our language, then, this conception is incoherent.

[11]To say that the meaning of a word is determined by its grammar *in general*, is to make philosophical use of "to determine". The use of the verb "to determine" is related to the verb "to fix", "to settle". Although language can be understood in these terms for certain purposes, the concept of language also suggests dynamism and change.

[12]It is important to say that, if some uses of language can be understood as not being determined based on rules, we do not mean that language use has no restrictions whatsoever. The use of concepts is not a matter of capriciousness. The use of some concepts, for instance, "pain", can be explained based on pre-linguistic human facts (PI, 244). See also Kuusela 2019, p. 184.

machine is philosophically problematic. On the other hand, Hacker's perspective rejects ascriptions of mental concepts to some objects and machines because the grammar of these concepts, according to Hacker, excludes ascriptions to creatures and objects that do not behave like humans, i.e. to some creatures and objects that cannot be said to have a mind. What both perspectives fail to see is that ascribing mental concepts to creatures and objects whose behavior is significantly different from human behavior does not necessarily presuppose the ascription of a mind. In other words, the ascription of mental concepts to machines does not mean that our "attitude towards [them] is an attitude towards a soul" (PPF iv, 22). The technical use of *learning* in computer science represents an example of this feature of the grammar of mental concepts.

We do not disagree with Bostrom in that it is possible to apply mental concepts to software. However, for them to be meaningful, these ascriptions must be embedded in a practice, for instance, computer science and the development of machine learning models. Thus, we cannot overlook the different uses of our mental concepts. Following Hacker, the use of mental concepts is somehow bounded. We cannot (or rather, we do not) seriously ascribe thinking to a chair (PI, 361). However, the perimeter of the grammatical space of our mental concepts is not sharply established and fixed by the sense-determining rules for their use, which in turn are related to the behavior of human beings and to creatures whose behavior is similar to human behavior.[13] In this sense, grammar does not exclude ascriptions of mental concepts to machine learning algorithms and some machines and objects that display a different behavior from humans. Despite the machine learning model's not having the repertoire of behavior that can be observed in humans, the grammar of mental concepts renders these ascriptions meaningful. See section 4.1.

5.3 A change in grammar: fluctuations in the use of *learning*

The last two sections moved slightly away from the analysis of machine learning and human learning, which represent our main object of investigation. However, the analysis of such instances of learning gives us an approximate idea of the status of other mental concepts that are applied to machines and software in science, particularly in the field of AI. Let us focus again on the concept of learning.

What can we say about the use of this concept in machine learning and computer science? First of all, our analysis has shown that the use of the concept of learning in computer science does not have the full semantic spectrum of the concept's original use in natural language (see section 2.1). This does not mean, however, that the technical use of the concept of learning is not legitimate. The concept of learning in computer science represents a new, actual use. A sufficient amount of computer scientists and experts coined the concept and started to use it in a particular way for it to find a new domain of application. To understand this new, technical concept, we described two of its uses in their specific context of application. Such analysis showed, for instance, (i) that the new, technical use of *learning* —so far— has not

[13] As stated earlier, we agree with Hacker in the sense that human behavior represents a strong criterion (but not the only one) for the ascription of mental concepts to some creatures, i.e. to creatures whose behavior resembles human behavior.

affected its ordinary use, (ii) that the application of *learning* concerning machines does not mean the same as its application concerning human beings (and animals that share some forms of human behavior), although the performance of the machine resembles, in a restricted way, human behavior, (iii) that the technical use of *learning* represents a secondary or derived application of its use in natural language (PI, 282; PPF xi, 275-8), (iv) that in some sense the grammar of *learning* is not something fixed, but allows variations in its use. In other words, the grammar of *learning* allows us to apply this concept beyond humans, based on the machine's structure, functioning, and rendering of results, (v) that the use of *learning* in computer science fulfills specific purposes, for example, to communicate the machine's performance in data analysis, to express that the scientists managed to obtain certain results, etc.

Another implication of our analysis is that the technical use of *learning* can be viewed as an *extension* and as a *fluctuation* of its ordinary use. Let me explain this point in more detail. Mapping the different uses of a family resemblance concept provides an overview of its general range of applications (PI, 122). Once we have surveyed the "logical geography" (Ryle 2009, p. 5) of the concept of learning, its technical use can be understood as an *extension* of its ordinary set of uses. The new area that the technical use of *learning* represents is characterized by sharing overlapping features with the ordinary use of this concept, and at the same time, by being significantly different in the way it is applied. By zooming in and focusing particularly on the technical use of *learning*, this concept represents a *fluctuation* concerning its ordinary use. With this new usage we witness a change or variation in the grammar of *learning*. When making use of the technical concept while operating the machine, the scientist does not have in mind the ordinary use of *learning*. This feature of the concept shows that the technical use is different enough to be considered a non-trivial, substantial fluctuation.[14]

Language is sometimes represented as consisting of strict sense-determining rules for the use of words to clarify certain philosophical problems. This view, however, does not manage to give an account of the fluctuations and changes in the grammar of certain concepts. To understand these matters, we need to remember that language is not a rigid, fixed feature on the margins of human practices. Rather:

> Language just is a phenomenon of human life. (RFM VI, 47)

> We are talking about the spatial and temporal phenomenon of language (PI, 108)

Language is embedded in and intertwined with our different ways of living. Thus, how we use concepts is not determined. New concepts are born, diverse uses are invented, different forms of representation emerge, concepts and language-games may become obsolete (PI, 23). Within its regularity and solidity, language presents anomalies, variations, and dynamism. This does not mean that language changes

[14]Klagge stresses the difference between the trivial and the important change in the use of a concept. The first refers to stipulation, and the second refers to a substantial change in the use (meaning) of the concept, which Klagge calls "evolution of concepts in time" (Klagge 2017, p. 196ff). It is not clear whether this classification can be sharply established. On the one hand, the change in the use of *learning* can be seen as a stipulation: scientists agreed that *learning* means x. On the other hand, the change in the use of *learning* can be seen as substantial. Scientists use this concept because the way the machine performs has criss-crossing similarities with the way humans learn. Perhaps these features would not be captured if scientists decided to use another concept.

uniformly. Some parts or areas of our language change faster than others. Think about the modifications of language and the use of our concepts fostered by political activism such as equality among genders, by scientific and technological inventions, such as airplanes, computers and the internet, by vegetarianism and animal rights movements, etc. As Klagge points out: "Just as one should not be an essentialist about the nature of concepts at a time, one should not be an essentialist about the nature of concepts over time" (Klagge 2017, p. 200-1). In this sense, language changes, and this is a proposition of grammar.

Another aspect of the dynamism of language is that some propositions that were considered to be grammatical are now regarded as empirical, and vice versa. In the first half of the twentieth century, for instance, "I have never been on the moon" was considered to be a grammatical proposition (OC, 111).[15] Having in mind this time and context, this expression does not give information about the world and does not convey an experience. This proposition is interesting because the current state of science and technology has allowed some humans to walk on the surface of the moon. Some propositions, indeed, modify their status. Nowadays "I have never been on the moon" can be considered to be, among a specific group of people, an empirical proposition. Similarly, "The machine cannot learn", expression that may have been regarded as grammatical before, now can be considered to be empirical. In some areas of computer science, an expert may use this expression pointing that the machine was not able to detect the patterns in a certain data set, or that the mathematical model in the algorithm is not appropriate, etc. Here we can see that the meaningful applications of mental concepts regarding machines and software are not made by overlooking the differences between their technical and ordinary usages. Rather, meaningful applications of mental concepts concerning machines are embedded in certain practices.

This investigation represents an attempt to shift our focus and attention. The problem is not whether a machine can think, learn, or understand in a strong sense. Rather, the question is in which sense it is possible to apply mental concepts to machines and software meaningfully.

[15]Strictly speaking, this is not a grammatical proposition in the sense that it does not represent a rule for the use of words. Nevertheless, this proposition has the role of a rule, in the sense that we presuppose it in certain language-games (OC, 95). This means that it works as a framework within which our empirical propositions interact. As Wittgenstein points out, some propositions are *fixed* for language to function, just like the hinges permit us to open and close a door (OC, 341 ff). These propositions belong to the scaffolding of our thoughts, they are non-analyzed statements, presupposed beliefs that are to be accepted for us to make judgments (OC, 211, 308).

Chapter 6

Concluding Remarks

1. **Summary and contribution**. The general aim of this investigation was —from a Wittgensteinian angle— to gain clarity about some of the uses of the concept of learning concerning machines and software as displayed in artificial intelligence. The second chapter presents an analysis of the concept of learning, which has not received enough attention in Wittgenstein scholarship, although it plays a central role in Wittgenstein's later thought (see Weber 2019 and Williams 1999a). Its analysis can help us to fathom in a new light sections of the *Investigations* that deal with problems of philosophy of mind and philosophy of language, such as concept formation, understanding, actions, and reactions, among others. The concept of learning shares features with other mental concepts and at the same time it contributes to the description of their use, when asking, for instance, how someone acquires a certain concept. In this chapter it was argued that *learning* is a family resemblance concept. On this basis, the use of *learning* in natural language was described through two language-games. The first one, color learning, represented a familiar instance of human learning. The second language-game, autodidacticism, was described as being a less familiar case of *learning*. Similarities and differences of these uses, and related concepts were offered in the final section.

The third chapter can be seen as a contribution to the philosophy of science, particularly to the philosophy of AI, in the sense that it offers a description of the concept of learning in machine learning employing two actual machine learning models, i.e. supervised and unsupervised. The reason for choosing these models was that their names are conceptually related to the familiar and less familiar instances of human learning described earlier, i.e. the situations in which children learn with the aid of a supervisor or expert and situations in which someone is said to learn "by herself", respectively. Although simple, these models remain relevant for our understanding of more complex algorithms used in science, for instance, neural networks in natural language processing systems. Interdisciplinary research is fundamental. In philosophy we would not gain substantial understanding by overinterpreting the meaning of our concepts and overlooking their different uses in a certain scientific practice. We need to analyze that practice and see how it works. It is in this sense that science is relevant to philosophy. Similarly, philosophical analysis can benefit science by clarifying the actual uses of concepts. As we tried to stress all along this investigation: scientific claims in AI, i.e. claims based on empirical research, are not the target of philosophical criticism. The role of philosophy, in this case, is to point to the incomprehension allocated in the conceptual framework of the scientist. Con-

ceptual clarity fosters scientific progress. From this point of view, scientists can gain a better understanding of their conceptual assumptions and thus of their practice (see Hark 1990, p. 270).

In chapter four, similarities and differences between machine and human learning were shown, by making use of some of Wittgenstein's remarks on mental concepts. The last section of this chapter represented graphically the differences we found using a table. This in turn helped to show that some mental concepts found a new, legitimate use in computer science (fifth chapter). This, it was argued, implies a significant difference in the meaning between the technical and ordinary concept of learning. Ascriptions of mental concepts to machines do not necessarily imply the ascription of a mind, something which is sometimes assumed in philosophical discussions about strong artificial intelligence. This point was made to show an incomprehension shared by two different perspectives —Bostrom and Hacker— about the meaningfulness of the application of mental concepts to software and hardware. The contribution of chapters four and five is two-folded. On the one hand, it was shown that computer scientists make use of mental expressions that —due to their vagueness, complexity, and family resemblance— may lead —if philosophically misinterpreted— to puzzlement about thinking machines and strong artificial intelligence. The concept of learning encompasses a great number of different features which are connected by nothing but family resemblances; a typical philosophical misunderstanding arises if we unjustly assume that all these features apply in all cases, or unjustly transfer some of the features of the learning instance A to the learning instance B —where they do not apply. On the other hand, the comparison between the technical and ordinary use of mental concepts also fosters our understanding of important features about the human mind, for instance, that human learning is related in a strong sense to the concept of action, i.e. the individual plays an active role in the process of learning a new ability.

It is worth noting that based on its structure and direction, this investigation can be regarded also as a case study of the notion of family resemblance, which is central to this work: by having shown some of the uses of the concept of learning in its familiar field, and then showing its expansion in computer science, we were able to trace important similarities and differences between both uses, an analysis that opened a view on the grammar of at least some of our mental concepts.

2. **Open problems and further lines of research**. Day by day technological advancements make AI machines and gadgets more accessible to us. The use of mental concepts concerning machines and software is thus increasing and becoming part of our everyday language. The emergence of virtual assistants, neural networks chess programs, etc. strongly encourage that agency, intentionality, thinking are ascribed to machines. In this sense, further mental concepts and problems concerning natural language processing, machine perception, computer vision, among others, deserve to be philosophically considered from the perspective that guided this work. In connection with this point, questions of philosophical methodology and the applicability of Wittgenstein's philosophy concerning the analysis of scientific practices remain open. This kind of theoretical investigation can be useful to elaborate on the practical implications of AI. The analysis of concepts like responsibility, data and personal information management, and the right to privacy can be addressed in the future based on the results partly obtained in this work. Additionally, it is important to consider the relationship between cognitive science and AI. Both areas

share important assumptions that need to be examined, for instance, the view that the human mind is a computer, i.e. an information processing system (functionalism) (see section 4.2 notes 4 and 6). These assumptions should not be accepted uncritically. Their analysis from a different angle is required. Lastly, the remarks on human learning show that, at least some of its aspects can nurture the discussion about *concept formation*, *primitive reactions* and *facts of nature* (see section 2.2 notes 7, 12, 13, and section 4.2 note 1), in the sense that the way humans learn could hint at an interpretation of Wittgenstein's philosophy based on these concepts.

Appendix A

Tools and Sources for the Machine Learning Algorithms

A.1 Tools and sources used to build the machine learning models

1. The programming language in which both machine learning algorithms were implemented is **Python**. For further information go to `https://www.python.org/`
2. The machine learning algorithms were built in **Google Colaboratory**, which is a Google cloud-based service that allows users to write, execute and share code written in **Python**. For further information go to `https://colab.research.google.com/notebooks/intro.ipynb`
3. **Scikit-learn** is a machine learning library for code written in **Python**, especially for supervised and unsupervised problems (Pedregosa et al. 2011). For further information go to `https://scikit-learn.org/`
4. **UCI Machine Learning Repository** is a website that collects different databases designed for machine learning algorithms. For more information go to `https://archive.ics.uci.edu/ml/index.php`
5. The **linear regression machine learning algorithm** (see subsection 3.2.1) was built with the *Student Performance Data Set* available at `https://archive.ics.uci.edu/ml/datasets/Student+Performance`. I followed the general instructions of the following tutorial `https://tinyurl.com/wqsr2bw`
6. The **clustering machine learning algorithm** (see subsection 3.2.2) was built by following the instructions of the following tutorial `https://tinyurl.com/u7rust7`. The data set was modified substantially. The original data set can be found at `https://tinyurl.com/s2byavb`

A.2 Sources of explanatory graphs

1. Figure 3.1: 'Support Vector Machine Graph' was obtained from **scikit-learn**. The data was modified.
   - Source: `https://tinyurl.com/wwdjjxb`
2. Figure 3.3: 'Linear Regression Graph' was obtained from **scikit-learn**. The data was modified.
   - Source: `https://tinyurl.com/uvbu6zx`
3. Figure 3.13: 'Overfitting' consists of three screenshots, obtained from Professor Alexander Ihler's YouTube channel. The name of the video is: 'Introduction (4): Complexity and Overfitting'
   - Source: `https://www.youtube.com/watch?v=VZuKBKd4ck4`
   - Alexander Ihler's website: `https://www.ics.uci.edu/~ihler/`
4. Figure 4.1: 'Overfitted Model' was obtained from **scikit-learn**.
   - Source: `https://tinyurl.com/s9jwcqx`

Bibliography

Alpaydin, Ethem (2014). *Introduction to Machine Learning*. 3rd. Cambridge, Massachusetts: MIT Press.

Aristotle (2004). *Nicomachean Ethics*. Ed. by Roger Crisp. Cambridge: Cambridge University Press.

Bennett, Maxwell and Peter Hacker (2003). *Philosophical Foundations of Neuroscience*. Oxford: Blackwell.

Bishop, Christopher (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Chalmers, David (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chance, Paul (2014). *Learning and Behavior*. 7th. Belmont, CA: Wadsworth.

Churchland, Paul (1981). "Eliminative Materialism and the Propositional Attitudes". In: *The Journal of Philosophy* 78.2, pp. 67–90.

Davidson, Donald (1963). "Actions, Reasons and Causes". In: *The Journal of Philosophy* 60.23, pp. 685–700.

Deutsch, Diana (2013). "Absolute Pitch". In: *The Psychology of Music*. Ed. by Diana Deutsch. 3rd. New York: Elsevier. Chap. 5, pp. 141–182.

Fodor, Jerry (1978). "Propositional Attitudes". In: *The Monist* 61.4, pp. 501–523. ISSN: 00269662.

Forster, Michael (2010). "Wittgenstein on Family Resemblance Concepts". In: *Wittgenstein's Philosophical Investigations: A Critical Guide*. Ed. by Arif Ahmed. Cambridge: Cambridge University Press. Chap. 4, pp. 66–87.

Foster, Michael (2018). *Aging Japan: Robots may have role in future of elder care*. URL: `https://www.reuters.com/article/us-japan-ageing-robots-widerimage-idUSKBN1H33AB`.

Glock, Hans-Johann (1996). *A Wittgenstein Dictionary*. Oxford: Blackwell.

— (2014). "Reasons for Action: Wittgensteinian and Davidsonian Perspectives in Historical and Meta-Philosophical Context". In: *Nordic Wittgenstein Review* 3.1, pp. 7–46.

Hacker, Peter (2009). "Philosophy: A Contribution, not to Human Knowledge, but to Human Understanding". In: *Royal Institute of Philosophy Supplement* 65, pp. 129–153.

— (2010). "The Development of Wittgenstein's Philosophy of Psychology". In: *Mind, Method and Morality: Essays in Honour of Anthony Kenny*. Ed. by P. Hacker and J. Cottingham. Oxford: Clarendon Press. Chap. 13, pp. 275–305.

— (2012a). "The Relevance of Wittgenstein's Philosophy of Psychology to the Psychological Sciences". In: *Deutsches Jahrbuch Philosophie* 3, pp. 205–223.

— (2012b). "Wittgenstein on Grammar, Theses and Dogmatism". In: *Philosophical Investigations* 35.1, pp. 1–17.

Hardwick, Charles (1971). *Language Learning in Wittgenstein's Later Philosophy*. The Hague: Mouton.

Hark, Michel Ter (1990). *Beyond the Inner and the Outer: Wittgenstein's Philosophy of Psychology*. London: Kluwer Academic Publishers.

Hertzberg, Lars (2011). "Very General Facts of Nature". In: *The Oxford Handbook of Wittgenstein*. Ed. by Oskari Kuusela and Marie McGinn. Oxford: Oxford University Press. Chap. 16, pp. 351–374.

IBM (2017). *IBM Watson Beat*. URL: `https://www.ibm.com/case-studies/ibm-watson-beat`.

Klagge, James (2017). "Wittgenstein, Science and the Evolution of Concepts". In: *Wittgenstein and Scientism*. Ed. by Jonathan Beale and Ian Kidd. London: Routledge. Chap. 11, pp. 193–208.

Kuusela, Oskari (2013). "Wittgenstein's Method of Conceptual Investigation and Concept Formation in Psychology". In: *A Wittgensteinian Perspective on the Use of Conceptual Analysis in Psychology*. Ed. by Timothy Rancine and Kathleen Slaney. New York: Palgrave Macmillan. Chap. 2, pp. 51–71.

— (2019). *Wittgenstein on Logic as the Method of Philosophy: Re-examining the Roots and Development of Analytic Philosophy*. Oxford: Oxford University Press.

Malcolm, Norman (1963). "Three Lectures on Memory". In: *Knowledge and Certainty: Essays and Lectures*. New Jersey: Prentice-Hall, pp. 185–240.

Mitchell, Tom (1997). *Machine Learning*. New York: McGraw-Hill.

Nelson, Katherine (2009). "Wittgenstein and Contemporary Theories of Word Learning". In: *New Ideas in Psychology* 27.2, pp. 275–287.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pichler, Alois (2018). "Wittgenstein on Understanding: Language, Calculus, and Practice". In: *Wittgenstein in the 1930s: Between the Tractatus and the Investigations*. Ed. by David Stern. Cambridge: Cambridge University Press. Chap. 2, pp. 45–60.

Ryle, Gilbert (2009). *The Concept of Mind*. New York: Routledge.

Schulte, Joachim (1993). *Experience and Expression: Wittgenstein's Philosophy of Psychology*. Oxford: Oxford University Press.

Searle, John (1980). "Minds, Brains, and Programs". In: *Behavioral and Brain Sciences* 3.3, pp. 417–424.

Shanker, Stuart (1998). *Wittgenstein's Remarks on the Foundations of AI*. London: Routledge.

Silver, David et al. (2017). "Mastering chess and shogi by self-play with a general reinforcement learning algorithm". In: *arXiv preprint arXiv:1712.01815*.

Tejedor, Chon (2017). "Scientism as a Threat to Science". In: *Wittgenstein and Scientism*. Ed. by Jonathan Beale and Ian Kidd. London: Routledge. Chap. 1, pp. 7–27.

Thagard, Paul (2012). "Cognitive Architectures". In: *The Cambridge Handbook of Cognitive Science*. Ed. by Keith Frankish and William Ramsey. Cambridge: Cambridge University Press. Chap. 3, pp. 50–70.

Tripodi, Paolo (2020). *Analytic Philosophy and the Later Wittgensteinian Tradition*. London: Palgrave Macmillan.

Turovsky, Barak (2016). *Found in translation: More accurate, fluent sentences in Google Translate*. URL: https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/.
Waismann, Friedrich (1968). *The Principles of Linguistic Philosophy*. London: Macmillan.
Walsh, Matthew M. and Marsha C. Lovett (2016). "The Cognitive Science Approach to Learning and Memory". In: *The Oxford Handbook of Cognitive Science*. Ed. by Susan E. F. Chipman. Oxford: Oxford University Press. Chap. 11, pp. 211–230.
Weber, Désirée (2019). "A Pedagogic Reading of Wittgenstein's Life and Later Works". In: *Journal of Philosophy of Education* 53.4, pp. 688–700.
Williams, Meredith (1999a). "The Philosophical Significance of Learning in the Later Wittgenstein". In: *Wittgenstein, Mind and Meaning: Toward a Social Conception of Mind*. New York: Routledge. Chap. 7, pp. 188–215.
— (1999b). *Wittgenstein, Mind and Meaning: Toward a Social Conception of Mind*. New York: Routledge.
Wittgenstein, Ludwig (1967). *Zettel*. Oxford: Blackwell.
— (1969). *Blue and Brown Books: Preliminary Studies for the 'Philosophical Investigations'*. Oxford: Blackwell.
— (1976). *Wittgenstein's Lectures on the Foundations of Mathematics, Cambridge 1939*. Ed. by Cora Diamond. Chicago: The University of Chicago Press.
— (1980a). *Remarks on the Philosophy of Psychology I*. Oxford: Blackwell.
— (1980b). *Remarks on the Philosophy of Psychology II*. Oxford: Blackwell.
— (1983). *Remarks on the Foundations of Mathematics*. 29. London: MIT Press, p. 368.
— (1991). *On Certainty*. Oxford: Blackwell.
— (1993). "Cause and Effect: Intuitive Awareness". In: *Philosophical Occasions*. Ed. by James Klagge and Alfred Nordmann. Cambridge: Hackett Publishing. Chap. 12, pp. 371–426.
— (2005). *The Big Typescript: TS 213*. Ed. by Grant Luckhardt and Maximilian Aue. Oxford: Blackwell.
— (2009a). *Philosophical Investigations*. 4th. Oxford: Wiley-Blackwell.
— (2009b). "Philosophy of Psychology: A Fragment". In: *Philosophical Investigations*. 4th. Oxford: Wiley-Blackwell, pp. 182–243.
Zhang, Chiyuan et al. (2016). "Understanding deep learning requires rethinking generalization". In: *arXiv preprint arXiv:1611.03530*.