4.0 Engineering and Human Values

Managing Inductive Risk in the Use of Big Data Analytics for the Predictive Maintenance of Railway Systems

4.0 Engineering and Human Values

Managing inductive risk in the Use of Big Data Analytics for the Predictive Maintenance of Railway Systems

Pablo Muñoz

Master Thesis

Supervised by dr. Koray Karaca

Examined by dr. Peter Steigmaier

MSc Philosophy of Science, Technology and Society - PSTS

University of Twente

Faulty of Behavioural, Management, and Social Sciences

Enschede, the Netherlands

August 2022

Acknowledgements

First and foremost, I want to thank dr. Koray Karaca for his dedicated and clear guidance and support. It helped me not to lose myself in this topic that required a robust and detailed analysis.

Also, I want to thank dr. Peter Steigmaier and dr. Karaca. The discussions we held and, overall, their courses were of great intellectual and professional value. In the same way as epiphanies, they resolved succinctly questions that wandered heavily in my mind, and revealed new horizons and perspectives of inquiry, reflection, and application.

Finally, I want to thank to my family and friends for their support, specially to my sister Mariana for her fabulous artistic contribution.

Thanks a lot!

Cover design by Mariana Muñoz

Contents

Summary

This research studies the use of Big Data Analytics technologies, in particular deep learning (DL) models, in the predictive maintenance (PdM) of railway systems. The specific aspect that it analyses is how inductive risk arises in this use. The inductive risk of DL is already studied by the philosophers of inductive risk and by the epistemologists of Machine Learning (ML) and DL. Also, the PdM of railway systems is a practice where inductive risk arises. In fact, (1) it is performed in systems that are safety-critical and, thus might impact human values; (2) it uses statistical analysis in its decision-making processes; (3) it requires the analysis of complex datasets, i.e., datasets that cannot be modelled by predetermined parameters, for making predictions and taking decisions. Therefore, this research aims to analyse in a synthetic way how the inductive risk assessment of Big Data Analytics can be used for assessing the inductive risk of the PdM of railway systems.

This research uses an empirical analysis, specifically a two cases study, for answering this question. As the inductive risk of the use of DL for the PdM of railway systems is situated specifically in the inspection activities of these maintenance practices, it studies two technological solutions where DL models are used for improving these inspection activities of these practices. In the first place, this research defines the general concepts that guide the assessment of inductive risk in the use of DL for the PdM of railway systems. In the second place, it uses these concepts for systematically studying the technical documentation where these solutions are described and explained. Thus, it interprets these descriptions and explanations, and compares and contrasts these two cases at the light of these concepts.

The major results of this research are two. In the first place, it reveals that the discussion of the inductive risk that arises in the use of DL for the PdM of railway systems is a discussion about how the processes for acquiring data that describes the physical status of these systems impact human values. In the second place, it shows that assessing inductive risk in this use consists in evaluating the possibilities and restrictions of managing the probabilities of inductive errors at the inspection activities of these maintenance practices.

At the theoretical level, this research focuses on identifying those aspects of the PdM of railway systems that define the specific subtopics of inductive risk that must be used for assessing this risk in these maintenance practices. Also, it shows how these subtopics are reformulated when these practices are mediated by DL models. At the empirical level, it uses the concepts produced

in the theoretical analysis of how inductive risk appears in the use of DL for the PdM of railway systems for determining the aspects of the technical literature that must be studied for understanding this inductive risk. Also, it transforms the theoretical questions into question about how to assess the multiple technological features of these solutions for giving answers to the questions generated by the theoretical discussion of this risk.

The major conclusions of this research are three: 1) The inductive risk that arises in the use of DL for the PdM of railway systems is a risk that must be managed at the three stages of the use of DL models: training, validating, and testing. Also, this management consists in identifying whether the contexts of these stages are context rich or poor in data. 2) This identification can be done by evaluating whether the inductive errors that might be produced in this use have equal costs or not. 3) Maintainers and other stakeholders involved in PdM practices can identify whether these inductive errors have equal costs by assessing how the inspection activities of these practices manage the probabilities of these errors through the combination of multiple data management methodologies, i.e., by the performance of design practices.

Chapter 1: Introduction

1. Technological Context

Industrial revolutions (Schiele et. al., 2021) are characterised as historical periods where technological change transforms the ways services are delivered and occupational activities are performed and socially organised. One specific feature of the actual fourth industrial revolution (4IR), or industry 4.0 is that Big Data in combination with machine learning (ML) and deep learning (DL) computation is used for enhancing the decision-making processes in management and professional activities. In the first place, Big Data is the technique of extracting, processing, and modelling complex data sets through computational and software technologies with the aim of finding patterns in these data that could be used for predictive purposes (Karaca, 2021). In the second place, ML computation is a paradigm of computer algorithmic models where the design of these models is not based solely on rules or programming procedures but on the data these models gather (Alpaydin, 2016). This integration of ML and Big Data is known as *Big Data Analytics* (Karaca, 2021).

Two fundamental aspects of the application of Big Data Analytics to occupational and management activities are *safety* and *risk* (Chenariyan Nakhaee et. al., 2019; Xu et. al., 2013). Many services require that the professionals, managers, technicians, and other workers who deliver them take decisions that could put the life and health of the users of these services in danger. For example, if the turbine of a commercial aircraft is not maintained or repaired, then this vehicle could present failures during operation and possibly cause harm to the passengers[1]. Thus, if ML and DL algorithms aid this decision-making, then the implementation of these technologies in the delivery of these services can raise safety and risk concerns.

Nowadays there is an increasing use of Big Data Analytics in the maintenance of railway systems (Davari et. al., 2021; Xie et. al., 2020; Le Nguyen et. al., 2020; Chenariyan Nakhaee et. al., 2019). In the first place, these maintenance procedures require the acquisition and modelling of complex data. In fact, railways are systems with many interrelated components and, also, they are exposed to *dynamic environmental conditions* and *continuous use*. In the second place, maintainers prefer predictive methods. Other strategies, such as reactive and preventive time-based

[1] Examples and technical characteristics that require long descriptions are presented as footnotes.

maintenance, are either dangerous or produce costs overruns (Davari et. al., 2021; Xie et. al., 2020; Chenariyan Nakhaee et. al., 2019). It is fundamental to indicate that railway systems are *safety critical systems* (SCS) *(*Davari et. al., 2021; Xie et. al., 2020; Le Nguyen et. al., 2020; Chenariyan Nakhaee et. al., 2019; Xu et. al., 2013). This means that their failure could cause catastrophic outcomes that generate huge human, economic, and environmental loses. A clear example of these catastrophic failures are derailments. Therefore, these systems must operate under the highest standards of safety and efficiency. ML and DL models can use these complex datasets for producing data patterns that accurately determine the functioning, anomalies, and remaining useful life (RUL) time, among other fundamental features of railway components, and transform them into clear decision procedures about when and how to maintain a component for preventing failure. These features made them ideal for attaining optimal levels of maintenance efficiency and safety. For this reason, the use of Big Data analytics in *predictive maintenance* (PdM) is a relevant topic from the historical perspective of the 4IR.

2. Philosophical Relevance

Predictive reasoning is based on *scientific induction*. It consists in the process of gathering statistical evidence for supporting the truth or falsehood of a hypothetical claim (Hempel, 1965). As the accuracy of the predictions made by railway maintainers determines whether *human values* such as safety, economic efficiency, environmental preservation, among others, are positively or negatively impacted by the railway systems, the predictive maintenance (PdM) of these systems is fundamental to the discussion about how scientific reasoning and processes impact human values. This discussion about the impact of the scientific inductive process on human or non-epistemic values is the topic known as *inductive risk*. Therefore, the PdM of railway systems is a context where inductive risk is present and can be philosophically studied.

The threshold that assures scientists that a body of statistical evidence is sufficiently strong and, thus, can be used for supporting the truth or falsehood of a hypothesis might vary (Hempel, 1965). In the scientific contexts where this threshold is subject to variation, scientists must formulate acceptance rules that justify a threshold choice and, thereby, whether the statistical evidence gathered is sufficient for supporting the truth or falsehood of this hypothesis. For this

reason, at these contexts, the outcomes of the scientific inductive process are four: 1) Scientists accept the truth of a statement based on sufficient statistical evidence (true positive); 2) scientists reject the truth of a statement based on sufficient evidence (true negative); 3) scientists accept the truth of a statement based on insufficient evidence (false positive); 4) scientists reject the truth of a statement based on insufficient evidence (false negative). In some of the contexts where scientific induction is risky, statistical thresholds and the rules that set them are subject to variation because each of these thresholds and rules is aimed to protect specific human values. Therefore, the set of thresholds depends on how the preservation of human values is articulated.

This is the case of the PdM of railway systems, as at this context values such as efficiency and safety guide how predictive maintenance practices are implemented and performed (see section 1). For this reason, the discussion of inductive risk at this maintenance context is a discussion about the *management* of this type of risk. This means that the topic of inductive risk at this context is about setting *decision rules* about which level of statistical evidence should be adequate for accepting or rejecting a hypothesis by analysing the impacts on human values of the outcomes of accepting or rejecting these hypotheses. Inductive risk is related to the philosophical topics of *decision theory* and *statistical decision making*.

Additionally, in some cases, managing this risk must not be done solely during the testing of hypotheses, i.e., during the acceptance or rejection of these hypotheses (Karaca, 2021; Ohnesorge, 2020; Wilholt, 2009; Douglas, 2000). Also, scientists must manage it during the design of the tests of these hypotheses and during the evaluation of the methodological decisions used for designing these tests. The reason for this is that, in these cases, scientists performing inductive reasoning processes that impact human values also employ statistical evidence for supporting their test designs choices and their methods for evaluating these test designs. As all these decisions produce a specific outcome attached to a specific impact on human values, inductive risk is present in all these types of decisions. Therefore, inductive risk must be managed at the *distinct stages* of the scientific inductive process.

This is the case of the PdM of railway systems. As indicated in the earlier section, these systems are SCS and, consequently, the decisions concerning the design and evaluation of the maintenance procedures performed at these systems impact human values as much as the performance of these procedures. Thus, in these systems statistical evidence is used at the multiple stages of the use of these systems (design, validation, and testing) and, because of this, inductive

risk is present and must be managed at all these moments. Hence, the discussion of the management of inductive risk in this context is philosophically related to the discussion of the connection between the multiple stages of the *scientific process*.

Furthermore, two distinct philosophical perspectives propose two different methods to manage inductive risk at some of these *value contexts,* i.e., at some of the contexts where scientists seek to protect or preserve human values (Ohnesorge, 2020; Wilholt, 2009). As mentioned before, at value contexts, scientists must decide the rule for setting a statistical threshold based on how the scientific process might impact this preservation of values. Consequently, the abovementioned perspectives differ on how to decide this rule. *Methodological conventionalism* claims that this rule must be negotiated by all the stakeholders that represent each of the values that might be impacted by the scientific inductive process and embodied in a methodological convention (Wilholt, 2009). *Permissive empiricism* claims that, in some cases, conventions might be flawed because these stakeholders do not have the data that describes adequately the relationship between this process and these impacts. For this reason, for this second perspective, in these cases, conventions must guide the management of inductive risk temporarily. When new statistical evidence or data that models the relationship between scientific induction and the preservation of human values is available, these conventions should be revised and adjusted at the light of this new evidence (Ohnesorge, 2020). Therefore, the distinction between these two perspectives (Ohnesorge, 2020; Wilholt, 2009) shows that the management of inductive risk consists in differentiating those contexts where there are data that describes the relationship between scientific induction and the preservation of values from those contexts where there are no such data. This shows that the discussion about the management of inductive risk is a discussion about how the *availability of data* specifies how scientists must manage inductive risk.

The PdM of railway systems is a context where the availability of data influences the decisions for managing inductive risk. As mentioned in the previous section, the environmental conditions and the continuous use of these systems produce data about these systems that is continually changing and, consequently, the availability of some of these data also changes; sometimes these data are available; sometimes they are unavailable. Therefore, this context is relevant for the philosophical discussion about the relationship between *data* and managing inductive risk.

3. Problem

With the advent of the 4IR, different philosophers have studied how inductive risk arises in different Big Data Analytics societal applications. Among these applications, there are banking (Karaca, 2021), urban governance (Jansen, 2021), and justice administration (Biddle, 2020). A common pattern between these studies is that they acknowledge that inductive risk arises at the three stages of the use of them (training, validating, and testing). They also affirm that assessing inductive risk in these applications should aim at establishing decision procedures that protect the values of the users of these applications. Furthermore, they claim that these decision procedures should be framed under the constrains and possibilities given by the mediation by Big Data of societal governance and social activities. Hence, this shows that the assessment of inductive risk in Big Data Analytics must be approached through the lenses of (1) decision theory, (2) the multiple stages of scientific processes, and (3) how the availability of data regulates the use of Big Data Analytics. This type of analysis has not yet been done in the application of Big Data Analytics to the PdM of railway systems and of safety-critical systems (SCS) in general. As exposed in the previous section, inductive risk is also present in these maintenance practices, even in those that do not use Big Data Analytics applications. Also, this section reveals that the inductive risk existing in these maintenance practises must be approached through the three abovementioned conceptual frameworks. For this reason, on the one hand, there is a common framework that assesses two distinct scientific contexts where inductive risk arises and, on the other, these two contexts converge in the use of Big Data Analytics for the PdM of railway systems. This fact allows me to affirm that a fundamental question to be asked is *how should the inductive risk assessment of Big Data analytics applications be used for assessing the inductive risk of the PdM of railway systems?*

Thesis: The inductive risk assessment of Big Data Analytics should transform the philosophical questions that arise in the assessment of inductive risk of the PdM of railways systems into design questions that must be responded by the engineers and maintainers of Big Data Analytics solutions for the PdM of railway systems with the aim of protecting the human values at stake in these maintenance practices.

For answering this research question and evaluating whether this thesis is true, it is necessary to answer the following sub-questions:

Sub-question 1: How inductive risk arises in the PdM practices performed in railway systems? This question allows me to identify the questions that appear when I use the philosophical framework of inductive risk for describing the PdM practices of these systems. Using this framework shows that inductive risk is a type of risk present in these practices. Therefore, with this framework I will account for how the methodologies these practices use for maintaining railway systems produce inductive risks.

Sub-question 2: How inductive risk is managed in the use of Big Data Analytics societal applications? By answering this question, I will produce the general framework that guides the assessment of inductive risk in all types of societal applications of Big Data Analytics. This means that, with this framework, I will be able to assess if it is possible to address the multiple inductive risks present in these distinct applications by using a common set of concepts and methods. Also, it will allow me to know which elements I should approach in the specific Big Data Analytics solutions designed for the PdM of railway systems if I aim to manage the inductive risk present in these solutions.

Sub-question 3: How the specific design characteristics of Big Data Analytics solutions for the PdM of railway systems influence the management of inductive risk in these maintenance practices? The answer to this question will show the limitations of applying a general framework for assessing the inductive risk of Big Data analytics to specific contexts where Big Data analytics solutions are applied. Consequently, I will be able to integrate the concepts of this framework with the specific problems related to inductive risk that arise in these solutions. This integration consists in formulating the technical questions present in these solutions regarding the impact of human values in a terminology that can be addressed by this framework. Therefore, I will show that the specific technical problems related to inductive risk are part of a general philosophical reflection about this type of risk and its relation to the impact of Big Data analytics in societal activities such as maintenance.

4. Structure of the research

In the *second chapter* I expose the methodological framework I use for answering the research question and sub-questions. As I employ an *empirical research approach* focussed on the analysis of technical literature, I describe and justify the sampling of this literature, the aspects that I examine in it, and the steps I follow in this analysis. In this way, I expose the connection between the methodological decisions and the research question and sub-questions.

The first sub-question is answered in the sections 1 to 4 of the *third chapter* of this thesis. I expose the topic of inductive risk and analyse how it appears in the conceptualisation of PdM. First, I expose each of the sub-topics of inductive risk that are relevant to the PdM of railway systems. Second, I expose the main methodological features of this type of PdM. Third, I show how each of the categories described by these methodologies are related to these sub-topics.

In the section 5 of the third chapter and in the *fourth*, *fifth* and *sixth* chapters, I answer the second and third questions. As I mention in the methodology chapter, the analysis of the technical writing and its relation to inductive risk and the societal applications of Big Data Analytics must be done at a general level and at a specific level. The reason for this is that the framework that connects the societal applications of Big Data Analytics and inductive risk is both used for shaping how inductive risk arises in these societal applications and for giving philosophical recommendations about how the users of these applications should manage this risk in order to protect their human or societal values. Furthermore, as inductive risk is present at the three principal stages of the use of these applications, training, validating, and testing, the answer to these questions must be done at these three stages. Therefore, in each of these chapters I address one of the stages (in the fourth chapter I address testing, in the fifth chapter I address validation, and in the sixth chapter I address training) and analyse the relation between inductive risk, PdM of railway systems, and the societal applications at the general level and at the specific level (in the section 5 of the third chapter I introduce the concepts for doing the general analysis). At the general level, I analyse how inductive risk appears at each of the stages of the use of Big Data Analytics societal applications and which are the principles that users should follow for managing this risk and protecting their human values. At the specific level, I analyse two Big Data Analytics solutions and assess which of the technical features they propose are relevant to the discussion of how to use the inductive risk assessment of the societal applications of Big Data Analytics for characterising this

type of risk in the Big Data Analytics applications used in the PdM of railway systems. Additionally, I evaluate how the philosophical principles employed in this risk assessment can be used for interpreting the technical solutions related to inductive risk as philosophical answers about how to manage this type of risk in PdM. As both solutions might vary regarding both how to characterise and how to manage inductive risk, my analysis, besides interpreting these two solutions, consist in comparing and contrasting them.

In the seventh chapter I present the conclusion of this thesis. I argue whether the thesis proposed is the adequate answer to the research question or not. Therefore, I expose which type of questions can be used by engineers and maintainers to assess and mitigate inductive risk in the use of Big Data Analytics for the PdM of railway systems and, hence, to protect the values of the users of these systems.

Chapter 2: Methodological Framework

The methodological framework I use is based on the method of qualitative document analysis proposed by Altheide and Schneider (2013). Therefore, I use the categories and steps proposed by this method for structuring and justifying the methodological decisions I chose. In this chapter, I expose my methodological framework according to the four categories Altheide and Schneider use for explaining how researchers that use qualitative document analyses should justify their methodological decisions.

1. Problematisation of the unit of analysis

I use an empirical method. For this reason, I analyse the maintenance practices of railway systems where Deep Learning (DL)[2] models are used. The source of data used for making this analysis is the *technical literature* or *technical writings* railway maintainers and engineers use for documenting and exposing these practices (Society for Technical Communication, 2021). Consequently, I use a *document analysis* methodology for approaching these practices.

Following Bowen (2009:31), I use a document analysis methodology for three reasons. In the first place, technical documents present a *stable* source of data. This means that my inquiry does not alter the answers I can get during the investigation of the selected object of research. Due to the high number of mathematical formulas and diagrams required for studying inductive risk and DL (Karaca, 2021; Ohnesorge, 2020; Wilholt, 2009; Levi, 1962), I consider that other qualitative research methods could not gather and study the abstract expressions of these formulas and diagrams in a stable way.

In the second place, technical documents are *exact* (Bowen, 2009:31). According to the Society of Technical Communicators (2021), technical documents communicate the details necessary for enabling a communication between distinct areas of expertise. In this case, these

[2] In the previous chapter I discuss the application of Big Data Analytics in the PdM of railway systems and its relation to inductive risk. In the philosophical discussion, the term "Big Data Analytics" is interchangeable with those of machine learning (ML) and deep learning (DL), as the philosophical inquiry about this technology is directed toward the epistemological features of ML and DL models, i.e., of algorithmic models used for the extraction of data from complex datasets (Karaca, 2021; Zednik, 2021). For this reason, from this point I refer to DL and inductive risk, and to DL and PdM.

areas of expertise are the philosophy of inductive risk, the epistemology of DL and Machine Learning (ML), computer science and artificial intelligence, and maintenance. For this reason, using technical documents facilitate the description and analysis of the common concepts, categories and terms found in these areas.

In the third place, technical documents provide an *ample coverage* (Bowen, 2009:31; Society of Technical Communicators, 2021). This means that they express both the details of technical solutions, but also show the general concepts, categories, models, terms, and symbols used in the technical areas under discussion (for example, Serradilla, et. al. 2022 give a general picture of the use of DL in PdM; Marino, et. al., 2007 propose a specific technical solution). Furthermore, technical literature provides the state of the art and the evolution of a specific technology (Serradilla, et. al. 2022; Davari et. al., 2021; Xie et. al., 2020; Le Nguyen et. al., 2020; Chenariyan Nakhaee et. al., 2019). In this way, due to my research interest of analysing a general framework, i.e., the relationship between inductive risk, DL, and PdM by studying concrete maintenance methodologies and technical solutions, a qualitative research approach that provides at all conceptual levels data about the research object is adequate for my research.

2. Constructing a protocol

The three research sub-questions show that the analysis of inductive risk in the use of DL models for the PdM of railway systems must be approached at three levels. First, it is necessary to evaluate how inductive risk arises in the *maintenance methodologies* of railways. For this reason, I analyse a review about the use of DL in PdM (Serradilla, et. Al., 2022) and two reviews (Xie et. al., 2020; Le Nguyen et. al., 2020) and two surveys (Davari et. al., 2021; Chenariyan Nakhaee et. al., 2019) about the *use of DL in the PdM of railway systems*. For analysing these documents, I, first, expose the philosophical framework of inductive risk. Second, I expose the main concepts found in these documents that, from my perspective, are related to this philosophical framework. Third, I make explicit and discuss this relation between these frameworks and these concepts. I do this analysis at chapter 3.

Second, inductive risk is also a topic researched philosophically in the literature of the *epistemology of ML and DL* (Karaca, 2021; Jansen, 2021; Biddle, 2020). In the analysis done at chapter 3 where inductive risk and PdM are related, I explain why the inductive risk that exists in PdM must be analysed according to the *three stages of scientific processes (external, planning, and internal)*. This analysis made at the three stages is also done by the epistemologists of ML and DL (Karaca, 2021). Consequently, as far as I analyse inductive risk in the PdM of railway systems that use ML and DL models, I must, in the first place, relate how inductive risk exists in PdM with how inductive risk exists in DL based on these three stages. In the second place, I must analyse how the specific characteristics of multiple DL solutions for the PdM of railway systems influence the way how the inductive risk of DL relates to the inductive risk of PdM. Thus, as this three-stages analysis is common to inductive risk in the PdM of railway systems and inductive risk in DL, at the chapters four, five, and six, I, first, analyse in a general way how the inductive risk of the PdM of railway systems is related to the inductive risk of DL according to each of the three stages (in chapter four I analyse the testing or external stage, in chapter five I analyse the validation or planning stage, and in chapter six I analyse the training or internal stage). Second, I analyse how two cases of *specific solutions of DL for the PdM of railway systems* materialise or make concrete this relationship between the inductive risk of the PdM of railway systems and the inductive risk of DL. As this materialisation varies, besides my analysis, I compare and contrast these two cases. This is the protocol that guides my research[3].

3. Themes and frames

I use two different methods for accessing the technical literature I chose. I apply the first method to the analysis of the general framework of the use of PdM methodologies for the maintenance of railway systems in relation to inductive risk. This method consists in utilising the concepts of the inductive risk framework presented by Hempel (1965), Douglas, (2000), Wilholt (2009), Ohnesorge (2020), and Karaca (2021) for analysing the technical literature where this general framework of railway systems PdM methodologies is presented. Thus, I search the main concepts of the inductive risk framework that appear in this technical literature and associate and discuss the

[3] The additional technical literature I use is solely employed for clarifying the terms and concepts used in the principal technical literature.

descriptions and conceptualisations of these concepts done by the authors of this technical literature with the descriptions and conceptualisations done by the authors of the inductive risk framework.

The second method has two steps. The first step consists in relating and discussing the concepts of inductive risk found in the technical literature that describes in a general way the methodologies for the PdM of railway systems with the concepts of the inductive risk that arises in the design of DL models. This relation and discussion allow me to identify the concepts that appear both in the discussion of the inductive risk of the PdM of railway systems and in the discussion of the inductive risk that arises in these design activities. The second step consists in searching for these common concepts in the technical literature that exposes the two DL models for the PdM of railway systems I chose. In this search, I assess how the specific factors of each of these solutions shape this relationship between the DL inductive risk and the railway systems PdM inductive risk. In this way, I analyse how the design decisions, definitions, and classifications made by the designers of these two models decide the specific features of how this relationship should be conceptualised and applied.

As I mentioned before, the analysis of how DL inductive risk and railway systems PdM inductive risk are related must be done at the three stages of the scientific process. For this reason, the application of the second method is distributed in each chapter according to the stage I analyse in each of these chapters. It is important to highlight that I have found a pattern in the two technical solutions I chose. Both have a section where they address how they trained the DL modelling solution (*description of the solution* chapters), a section where they expose the contextual constrains that specify and, consequently, validate these design decisions (*introduction*), and a section where they test the solution and discuss the outcomes of this test (*experimental results*) (Santur, Kakaröse, and Akin, 2017; Marino et. al., 2007). For this reason, each of these sections is analysed at the chapter where I analyse each of the stages of the use of DL models and of scientific processes (testing/external; validating/planning; training/internal). In this way, the application of these methods and the order of their application are the frames that organise how the themes of this research are approached.

4. Theoretical samplings

For selecting the sample of my research, I apply one sampling reasoning for selecting the documents required for the general study of railways systems PdM methodologies and a second one for selecting the cases that expose specific DL solutions. Regarding the first sampling reasoning, I use the general review of the application of DL in PdM done by Serradilla, et. al. (2022) for identifying the main concepts that are common to all PdM practices. These concepts allow me to understand in a more accurate way how DL is applied in the specific sector of railway systems maintenance. Also, I use these four reviews and surveys of the state of the art of the use of DL in the PdM of railway systems (Davari et. al., 2021; Le Nguyen et. al., 2020; Xie et. al., 2020; Chenariyan Nakhaee et. al., 2019) because they are the most recently published documents according to the search that I did. Furthermore, four of these documents were sufficient for extracting the common concepts of the use of DL in the PdM of railways systems.

Following Zainal's (2007) view that claims that using multiple cases allows researchers to generalise their findings, I chose to study two cases where DL models are applied for the PdM of railway systems. Thus, I use these two cases for finding the general patterns that exist in this application. Also, following Zainal (2007), I use multiple cases for finding the differences between the cases regarding how they conceptualise and make concrete this application and, consequently, for understanding which are the limitations of these general patterns. Therefore, the analysis of these cases consists in comparing and finding the similarities between the two and in contrasting them and finding the differences. I limited the study to two cases for complying with the research length requirements. The reason for this is that, as I analyse the cases at the three stages of the use of DL models, I make a deep and comprehensive study of these cases. This is the second reasoning that I use for sampling the technical solutions.

The first case that I chose is a digital camera inspection system used for improving the detection of missing fastening bolts in railways (Marino et. al. 2007)[4]. This system consists in an acquisition system mounted in a diagnostic train that films the railways. The DL model by means

[4] Before using computer vision systems (CVS) for inspecting tracks, maintainers did it manually. The inspector walked along the track and through visual inspection detected anomalies, in this case, missing fastening bolts. However, this process was unacceptably slow and lacked objectivity. Therefore, the decisions and predictions about when maintainers should apply a maintenance action were not taken and performed by them in the best way possible and produced an unacceptable rate of errors. Thus, using CVS for inspection reduces this error rate to the minimum (0.4%) and optimise the maintenance decision-making process.

of visual signal processing and real time visual inspection detects whether a bolt is positioned correctly or not. Missing bolts can cause failures in the rail track and the development of critical situations. For this reason, the system alerts railway maintainers that the tracks need maintenance. Therefore, the system aid maintainers in deciding and predicting when to perform a maintenance action in the track for avoiding critical situations. For more technical details of the system see Marino et. al. (2007).

The second case that I chose is a laser camera system used for inspecting rail tracks and detecting faults in them (Santur, Kakaröse, and Akin, 2017)[5]. As the other case, the system is mounted in a diagnostic train. This system combines a normal camera that captures rgb colour signals and a laser camera that, through triangulation, analyses the deepness of the rail surfaces. While the normal camera makes a visual representation of the system, the laser camera measures the three-dimensional embeddedness of the rail surfaces. Abnormal changes in these surfaces are signals of faults. The normal camera avoids that the laser camera registers structural changes as abnormal, by constructing the three-dimensional profile of the rail. The data captured in this inspection is used by the DL model for classifying the rails as faulty or healthy. With this information, maintainers can decide whether to maintain a railway or not. Other technical details are specified at the paper written by Santur, Kakärose and Akin (2017) where their proposal is explained.

Until this point, I have not indicated yet the difference between how to manage inductive risk in the PdM of railway systems and how to manage this type of risk in this type of maintenance applied to other systems that are also safety-critical, deployed in value contexts and affected by dynamic environmental conditions and continues use (I call them *intensive-maintenance systems*). The fundamental distinction is that these three characteristics of intensive-maintenance systems are specifically related to *inspections* in the case of railway systems[6]. By contrast, in other systems

[5] The purpose of the design of this system is also optimising inspection and, hence, aiding maintainers in their predictions and decisions about when to maintain a rail track. Nonetheless, the designers of this second system consider that using solely images is problematic, as oil and dust residues can be misclassified as faults and, hence, rails that are healthy, i.e., that do not have faults, can be classified as faulty. For this reason, they complement the visual inspection with a laser system. Therefore, both cases are inspection systems that use CVS for detecting, diagnosing and prognosing faults in rail tracks. The main difference is that, while the first one does not use a laser system, the second does. These inspection systems belong to the group of inspection systems for rail infrastructure (Santur, Kakaröse, and Akin, 2017; Marino et. al., 2007). The other group of inspections systems is the one aimed at inspecting vehicles (Davari et. al., 2021:10-12).

[6] In the PdM of railway systems, inspections are the activities by which maintainers acquire the data necessary for evaluating whether a system component is having an abnormal behaviour or presents abnormal characteristics, diagnosing which type of failure it might have due to these anomalies, and prognosing when it should be maintained

they might not be exclusively related to this type of maintenance activities[7]. Therefore, since I aim to study how to manage inductive risk in railway systems, I must focus my research on inspection activities[8]. For this reason, I selected two cases of DL models applied in the inspection activities of the PdM practices of railway systems.

to avoid this failure (Davari et. al., 2021; Xie et. al., 2020; Chenariyan Nakhaee et. al., 2019). The method for performing these activities decides whether the railway systems PdM practices satisfy, at the same time, the standards of safety and efficiency (Marino et. al., 2007). For this reason, safety and efficiency are the values that guide the method for performing these activities in these systems. Also, these activities are safety critical, since ignoring a faulty component due to a mistaken inspection might lead to a catastrophic consequence (Ghofrani et. al., 2018). Finally, the dynamic environmental conditions and continuous use of the railway systems might influence how well these activities are performed. As Santur, Kakaröse and Akin (2017) show, oil and dust can appear as structural failures of a rail track. Therefore, the impact of the dynamic environment and continuous use in the PdM of railway systems occurs in its inspection activities. These facts shows that the PdM context, i.e., the context of intensive-maintenance systems, that specifies the type of inductive risk that is studied in this thesis is defined by the fundamental characteristics of the inspection activities performed in this context. This is the reason why inspection must be the central activity to be studied if the objective of the thesis is to study the inductive risk that arises in the PdM of railway systems.

[7] Inspection activities, as other PdM activities, are classified according to the source from where the data necessary for making predictions are acquired. In the case of inspection, the source is the current physical state of a component of the system. Other data sources in PdM are the history of the component, the design data represented in the technical drawings of the system, the current and past operations performed in the system, reports about accidents and failures, among others (Ghofrani, et. al., 2018:232-233). Thus, these other sources configure other types of PdM activities. For example, when maintenance engineers combine the data produced in inspection activities with asset registration data (functional location, year of installation, technical details, etc.) they produce the notification datasets. Therefore, the sources of these datasets are two: the physical state and the registration procedure. Bukhsh, Saaed, and Stipanovic (2018) state that by analysing these notification datasets is also possible to predict when a railway component requires maintenance. Hence, *notification analysis* is another type of PdM activity.

[8] I cannot state whether in other intensive-maintenance systems the three fundamental characteristics that specify the type of inductive risk that arises at these PdM contexts converge in their inspection activities, or, perhaps, in other PdM activities. This would require a task that goes beyond the scope of this thesis: assess one by one these other intensive-maintenance systems and their PdM activities. However, I have demonstrated that, at least in the case of the PdM of railway systems, these three fundamental characteristics converge in inspection activities and, thereby, why, if I aim to study the inductive risk of these systems, I must approach these activities specifically.

Chapter 3. Predictive Maintenance of Railway Systems and Inductive Risk - a general framework

In this chapter I expose the aspects of the inductive risk problem that can be used for analysing this type of risk in the PdM of railway systems. After this, I show how these aspects are also used for studying risk in societal applications of DL. The concepts produced by relating at a general level these three topics, inductive risk, PdM, and DL societal applications, structure the theoretical framework I use for analysing the two technical solutions at chapters four, five, and six.

1. The inductive risk problem in predictive maintenance

A) The inductive risk problem

The *inductive risk problem* is a problem deeply discussed in the *philosophy of science* (Karaca, 2021; Ohnesorge, 2020; Wilholt, 2009; Douglas, 2000; Hempel, 1965; Jeffrey, 1956; Rudner, 1953). The term "inductive risk" was introduced by Carl Gustav Hempel (1965) in his work *Science and Human Values*. However, Richard Rudner (1953) was the first philosopher to discuss it. The fundamental question of this problem is how the inductive reasoning of scientists is related to the preservation and production of human values. This means that the discussion about inductive risk aims to investigate how this type of scientific reasoning is risky, i.e., can negatively affect human and societal values.

Inductive reasoning is understood by philosophers of inductive risk as the process of using *statistical evidence* for supporting a *scientific decision*. Some of these scientific decisions are the acceptance or rejection of hypotheses, the choice of methodologies, and how to gather, characterise, and interpret data (Karaca, 2021; Ohnesorge, 2020; Wilholt, 2009; Douglas, 2000; Churchman, 1948). The employment of statistical evidence is *inductive*, as a statement can be validly claimed as a true statement if, and only if, most of the cases show that this statement is true[9]. Therefore, inductive reasoning consists in identifying whether the outcome of a decision

[9] For instance, a group of scientists must choose between accepting a scientific hypothesis or rejecting it. For taking this decision, these scientists must gather statistical evidence that can be employed for supporting either the acceptance or the rejection of this hypothesis. Suppose that this hypothesis claims that the molecule bisphenol-A (BPA) is

course is true in most of the cases and, therefore, that this decision course must be taken by scientists.

B) Inductive risk in the predictive maintenance of railway systems

In the same way as the maintenance of other systems, the maintenance of railway systems is an activity that has three goals (Serradilla et. al., 2022; Hu and Dai, 2021). The first goal is improving the components of the system. This type of maintenance is called *improvement maintenance*. The second goal consists in mitigating or eliminating in advance an abnormal behaviour of a component that could lead to a failure of the system. Thus, the goal is to avoid that a failure occurs. This type of maintenance is known as *preventive maintenance*. The third goal is repairing a component that has already failed; in other words, is restoring the function of the system. The name of this third type is *corrective maintenance*.

As mentioned in the first chapter, the industry 4.0 made possible the use of Big Data Analytics in multiple professional activities. In the case of the maintenance of railway systems, this revolution improved preventive maintenance activities (Davari et. al., 2021; Xie et. al., 2020; Le Nguyen et. al., 2020; Chenariyan Nakhaee et. al., 2019). Before this revolution, preventive maintenance activities had to be managed in a timely manner. This means that they must be performed at regular time intervals. The reason for this is that maintainers could not capture and process all the complex data necessary for predicting when a component needed to be maintained. However, this produced costs overruns, as, on many occasions, components were maintained even if they did not present any abnormal behaviour. As mentioned in the first chapter, Big Data

carcinogenic. BPA is a molecule used in the production of polycarbonate plastics. As these plastics are used in many everyday products, the fact that this molecule can produce harmful effects such as cancer is relevant from a scientific and regulatory point of view. For more details about the philosophical relevance of how scientists and regulatory agencies have approached these dangers see Ohnesorge (2020), Wilholt, (2009), and Douglas (2000). Thus, these scientists must construct an experiment where they expose laboratory rats to this molecule. If the bodies of these rats start to develop cancerous characteristics, e.g., they start to develop tumours, then these scientists can use this evidence to support their decision of claiming that their hypothesis is true. This evidence is statistical because scientists cannot use a sole case in their experiment. They must expose multiple rats to BPA and see how many of them develop carcinogenic characteristics. If most of the rats develop these characteristics, they can validly use this evidence for supporting the acceptance of the hypothesis. If only a few numbers of rats develop them, they cannot validly employ this evidence.

Analytics is a technology used for analysing complex data sets and made predictions based on the statistical evidence gathered in these analyses. Therefore, this technology, in the case of the maintenance of railway systems, helps maintainers to predict or forecast when to maintain a component and, consequently, to shift from a *time-based* approach to a *condition-based* one (Davari et. al., 2021; Xie et. al., 2020; Le Nguyen et. al., 2020; Chenariyan Nakhaee et. al., 2019). In other words, maintainers, with the aid of Big Data Analytics, can set conditions about when to maintain a component based on a statistical analysis of the complex data produced by railway systems; in this way, there is no more the need to perform maintenance in a cyclical manner. This use of Big Data analytics in railway maintenance frames this type of maintenance as a problem about how to apply inductive reasoning. In fact, the decision about whether to maintain a component is determined by statistical evidence.

In the first chapter I have also mentioned that railway systems are Safety Critical Systems (SCS). This means that a failure in one of their components can affect negatively human values. For instance, a missing bolt in a rail track can cause a derailment, which, at the same time, can cause the death of passengers. In this case, the value of safe use of transportation systems is negatively impacted by the failure of the railway system. For this reason, railway systems are vulnerable to inductive risk. In the first place, their maintenance is based on statistical analysis and the use of statistical evidence. In the second place, their failure or malfunction can lead to negative societal consequences.

2. Managing inductive risk

A) Inductive risk and acceptance rules

I have claimed that evidence can be used for supporting a scientific decision if, and only if, most of the cases that compose this body of evidence show that the statement is true. However, a question must be posed: how many cases would be *most of the cases*? If 100 rats were exposed to BPA, how many of them must exhibit carcinogenic characteristics if the scientists want to claim that the hypothesis that this molecule is carcinogenic is true? 60 rats? Or 51? Or 80? As there are multiple answers to this question, the use of statistical evidence shows that this type of evidence cannot

support a scientific decision *conclusively* (Karaca, 2021; Ohnesorge, 2020; Wilholt, 2009; Hempel, 1965:90-93; Douglas, 2000). This means that additional elements must be involved in scientific decision-making for determining this statistical threshold that determines that most of the cases of a body of statistical evidence support this decision.

Hempel (1965:90-93) claims that these additional elements are known as *acceptance rules*. Thus, an acceptance rule determines which is the threshold necessary for claiming that most of the cases of a body of evidence support a scientific decision. Hempel observes that, if there are rules that determine these thresholds, scientific decision-making have four outcomes. 1) Scientists follow a decision or action course based on sufficient evidence (true positive); 2) scientists do not follow an action course based on sufficient evidence (true negative); 3) scientists follow an action course based on insufficient evidence (false positive); 4) scientists do not follow an action course based on insufficient evidence (false negative)[10].

Also, Hempel (1965:90-93) observes that the four outcomes of a scientific decision-making process must be classified into two groups according to how they impact a human value. The reason for this is that, as the possible outcomes of a decision related to an impact on a human value are just two, either this decision impacts the value or not, then the four outcomes of scientific decision-making must be classified in outcomes that cause this impact and outcomes that do not[11].This shows that the *management of inductive risk* is done in two steps: (1) scientists identify which outcomes might negatively impact a human value or group of human values. (2) Based on this identification, scientists formulate the acceptance rules that reduce the probability of producing

[10] Following the previous example, a true positive would be claiming that BPA is carcinogenic based on observing carcinogenic effects in a number of rats that surpasses the threshold; if this threshold is 80 rats, then the scientists claim that BPA is carcinogenic because they observe carcinogenic characteristics in 84 rats. A true negative would be rejecting that BPA is carcinogenic, as just 50 rats presented carcinogenic characteristics. A false positive would be accepting that BPA is carcinogenic, even if only 50 rats presented these characteristics. A false negative would be rejecting that BPA is carcinogenic, even if more than 80 rats presented these characteristics.

[11] Following the case of the carcinogenicity of BPA, claiming that BPA is carcinogenic based on insufficient evidence (false positive) and rejecting that BPA is carcinogenic, even if there is sufficient evidence to prove this carcinogenicity (false negative) would put at risk the safe use of this molecule. BPA is a molecule used in the production of polycarbonate plastics. Thus, claiming that BPA is carcinogenic without sufficient evidence will create fears in the use of these plastics. Therefore, all the consumers that need to use these plastics would be negatively affected. Also, rejecting that BPA is carcinogenic, even if there is sufficient evidence that shows the opposite, would put at risk consumers. They could be using a product that might cause them cancerous diseases. Consequently, while the true positives and true negatives will promote the safe use of polycarbonate plastics, false positives and false negatives will put users at risk. In other words, whereas the former two results impact positively the values of safety and health, the latter impact them negatively. Hence, inductive risk is present in these two latter outcomes.

these undesired outcomes by establishing the thresholds that reduce the number of bodies of statistical evidence that these scientists are allowed to use for supporting these outcomes.

B) Managing inductive risk in the predictive maintenance of railway systems

Despite the fact that Big Data Analytics facilitates the performance of predictive or condition-based maintenance strategies in railway systems, corrective and preventive time-based maintenance are still necessary in these systems (Le Nguyen et. al., 2020:21-26). On the one hand, corrective maintenance is necessary because some failures are impossible to predict or to model. For example, the death of an animal in the rail track is impossible to predict. On the other hand, preventive time-based maintenance is necessary because the data necessary for predicting some failures are either impossible or too costly to acquire. For instance, the data necessary for predicting when the paint of a train should be replaced. Therefore, two conditions must be met for applying predictive maintenance in railway systems (Davari et. al., 2021:3-5). 1) It should be possible to predict when the component will fail, i.e., the *remaining useful life* (RUL). 2) It should be possible to detect whether a component is having abnormal behaviour that would reduce its RUL. Whereas the first condition determines whether a failure is predictable, the second establishes if there are data that can be detected for predicting this failure[12].

This two-conditions criterion for applying PdM in railway systems shows that managing inductive risk in the context of railway maintenance is also a process of classifying four results into two groups. When deciding whether to apply PdM to a component, maintainers must establish, in the first place, whether the failures they aim to predict are predictable. In the second place, they must determine whether the data necessary for making this prediction can be acquired. If they confirm that a failure is predictable and successfully predict it through a statistical analysis, they will produce a true positive. The reason for this is that they are employing a data analysis

[12] For instance, predicting that the paint of a train will degrade and, hence, when this train requires a new layer of paint is possible. The problem is that acquiring the data for making this prediction has a very high cost. Consequently, the degradation of the paint of a train is an event that can be predicted but it is not possible to acquire the data for making such prediction. In other words, it meets the first condition, but not the second one. Instead, predicting than an animal will fall death in the middle of the rail track is very hard to predict, as this event has a very low chance of occurring for several reasons. Therefore, this second event does not meet neither the first nor the second conditions for being subject to PdM practices.

mechanism, i.e., an acceptance rule, adequately and, additionally, they are predicting a truly potential failure. If they deny that the failure is predictable and use a data analysis mechanism for showing that this failure is unpredictable, they will produce a true negative. In fact, they are adequately employing the data analysis mechanism for showing the impossibility of predicting a supposed failure. Correspondingly, if they make a prediction based on insufficient data or on wrongly applying a data analysis mechanism, they will produce a false positive; also, if they deny that a failure is predictable based on an inadequate data analysis, they will produce a false negative. Furthermore, correctly classifying a failure as unpredictable, will decrease the expenditure in technologies for predicting it and, simultaneously, will increase the expenditure in time-based and reactive maintenance. Also, if they predict a failure based on insufficient data, maintainers might perform a maintenance procedure when it is not necessary, or they might omit performing this procedure when it is necessary. In the first case, they will produce costs overruns and, in the second, they might put at risk the functioning of the system and, consequently, the safety of the users and the economic revenues produce by it. Thus, while true positives and true negatives will properly distribute the maintenance budget in the different strategies, as they truly state when PdM is possible and when not, false positives and false negatives will obstruct this cost optimisation procedure. Consequently, these two latter results will negatively impact the value of cost efficiency (besides impacting the value of safety; an impact produced by operating a system that is not properly maintained).

3. Managing global inductive risk

A) Inductive risk is present in the entire scientific process

Hempel (1965:90-93) claims that inductive risk is a type of risk that must be managed in the *testing* of scientific hypotheses. This means that scientists must, first, identify which acceptances or rejections of hypotheses impact human values. Second, based on this identification, they must formulate acceptance rules that reduce the probability of producing these acceptances or rejections by establishing the thresholds that reduce the number of bodies of statistical evidence that are allowed to be used for supporting these acceptances and rejections. However, other philosophers

of inductive risk (Karaca, 2021; Ohnesorge, 2020; Wilholt, 2009; Douglas, 2000) have shown that, first, other scientific decision-making processes, besides the testing of hypotheses, use statistical evidence for supporting the choices made by scientists. Second, the outcomes of these other processes might also impact human values. Therefore, in these other processes inductive risk is also present and must be managed.

Heather Douglas (2000) classifies scientific activities into two main categories: the scientific activities that are done at the *internal stage* of the scientific process and the scientific activities that are done at the *external stage*. The internal stage is the stage where scientists produce statistical evidence. Among the activities performed by scientists at these stages there are the choice of statistical methodologies, gathering and characterising data, and interpreting data. The external stage is the stage where scientists test hypotheses, i.e., where they establish whether the produced statistical evidence can be used successfully for proving the truth or falsehood of a hypothesis. Douglas claims that in the activities performed at the internal stage of science there is also inductive risk. The reason for this is that the methodological choices scientists follow at these stages determine whether there is sufficient statistical evidence for supporting a hypothesis[13].

Koray Karaca (2021), Miguel Ohnesorge (2020), and Torsten Wilholt (2009) state that there is a third stage where inductive risk is present. In fact, as the methodological decisions scientists take are constrained by the results they want to produce, scientists must assess not just how their decisions will affect the results of the scientific process, but also what happens if these beliefs about how a methodological decision produce a determinate outcome are false. Scientists can increase the probability of producing a result[14]. For this reason, scientists must acknowledge that the

[13] Following the case of the test of the carcinogenicity of BPA, Douglas (2000:569-572) claims that scientists working for the producers of polycarbonate plastics chose a strain of rat that is insensitive to oestrogens. BPA acts as an oestrogen since it can disrupt endocrine activity. As these disruptions are linked to the appearance of carcinogenic characteristics, choosing oestrogens-insensitive rats will decrease the probabilities of finding statistical evidence that supports the hypothesis that claims that BPA is carcinogenic. Thus, this methodological choice influences the results of the testing procedure. Since rejecting that BPA is carcinogenic and this rejection is in part based on a methodological choice, inductive risk is also present at the internal stages of science; choosing oestrogen-insensitive rats can affect the safety consumption of polycarbonate plastics.

[14] For example, they can use oestrogen-sensitive rats and increase the probability of finding that BPA is carcinogenic. However, they are not in complete control of this process of increasing the probability of this result. Suppose that BPA is carcinogenic just at a specific dose. Suppose also that there are no previous tests that register the connection between the dose of BPA and its carcinogenic effects. Therefore, even if scientists use oestrogen-sensitive rats, if they do not expose these test subjects to the adequate dose of BPA, they will not be able to increase the probability of finding that BPA is carcinogenic. As they do not know this because there is no prior empirical information about the relation between the dose and carcinogenic effects, then scientists will fail to increase the probability of their desired result: showing that BPA is carcinogenic.

connection between a methodological decision and a specific outcome is just probable and depends also on factors that they ignore. Consequently, they must consider the impact of following a methodological decision with the aim of producing a specific result in a situation where this connection between this decision and this result might not be true at all. As this consideration must be done before scientists decide how to perform a test, I call this stage where this consideration is made the *planning stage* of science.

B) Managing global inductive risk in the predictive maintenance of railway systems

PdM in railway systems, as in other systems, is a process that has four main stages. These stages are (1) anomaly detection, (2) failure diagnosis, (3) degradation prognosis and (4) mitigation (Serradilla et. al., 2022). *Anomaly detection* consists in evaluating whether the conditions under which a component is working are the normal conditions and whether this component is working correctly under these conditions. If the conditions under which the component works are not normal, the maintainer will register an anomaly. Also, if the conditions are normal, but the component does not work correctly, the maintainer will register an anomaly. For example, maintainers observe that part of a rail track is scratched. Therefore, they must determine whether these scratches should be considered signs of a future failure.

Failure diagnosis (Serradilla et. al., 2022) aims to determine whether the anomaly found in the previous stage will evolve into a failure or not. This is done by analysing the different combinations between the behaviour of the component and its working conditions, and by determining whether some of these combinations will produce a failure in the future. Following the previous example, maintainers evaluate whether using the rail 25 times per day (working condition) would make deeper the scratches (behaviour of the component) until a point where the rail breaks.

The *degradation prognosis* (Serradilla et. al., 2022) consists in modelling how, under the abnormal conditions or abnormal working, the damage of the component will evolve until it causes a failure. This prognosis is done by introducing the data gathered in the first stage in a mathematical model that represents the degradation evolution of the component and observing which are the outcomes: will the component fail under the present conditions? And, therefore, does it need

maintenance? For instance, maintainers introduce all the data from the scratched rail in a model of rail degradation and observe when, according to this model, the rail will break. Therefore, with the outcome of this model, maintainers can decide whether they need to maintain the component.

Mitigation (Serradilla et. al., 2022) is the subprocess of determining, based on the information produced by the three previous stages, which actions ought to be performed to avoid the failure of the component. Since maintainers already have all the information that allows them to determine whether the component needs maintenance or not, the mitigation process consists in designing a maintenance procedure in the case the component needs maintenance. Maintainers know that the scratched rail will fail in two months. Therefore, they must plan how to maintain the rail.

These four stages of the PdM process can be classified into the three stages of the scientific process, internal, external, and planning. In stage 1 (anomaly detection) maintainers evaluate whether the behaviour of a component is normal and whether the conditions under which it is working can be linked to a potential failure and, consequently, to the statement that this component needs to be maintained. Thus, as maintainers link possible interpretations of the working conditions and behaviour of the component with the truth or falsehood of hypotheses about whether to maintain it or not, this stage can be classified as a *planning stage*. In stage 2 (failure diagnosis), maintainers use the observations done in the first stage to evaluate whether the data gathered in these observations can confirm that the component needs to be maintained. As they analyse in this second stage the different combinations between the working conditions of the component and its behaviour or status, they evaluate how these different units of evidence can be articulated for claiming that a component needs to be maintained or for claiming that it does not require maintenance. For this reason, this stage can be classified as the *internal stage*. In stage 3 (degradation prognosis), maintainers by means of degradation models test whether the gathered data can be used for stating that the component will fail and needs maintenance. Since the outcome of this model gives the answer whether to maintain the component or not, it is the instrument utilised by maintainers to test their hypotheses about the maintenance of a component. For this reason, this third stage can be classified as the *external stage* of the PdM process. As in stage 4 (mitigation) maintainers already know whether to maintain a component or not, this stage does not need to be classified in the three stages of scientific processes.

4. Inductive risk, scientific modelling, and methodological paradigms

A) Inductive risk data and methodological paradigms

Karaca (2021), Ohnesorge (2020), and Wilholt (2009) claim that scientists must construct a *mathematical model* for mapping how a scientific decision impacts a human value and deciding the threshold for classifying a body of statistical evidence as adequate for supporting this decision. The reason for this is that a value is composed of sub-values and, consequently, scientists must evaluate how the impact on one of these sub-values affects the impact on the others. Therefore, if the negative impact on one of these sub-values entails the positive impact on other, scientists must establish a threshold that minimizes the impact on both, even if that means reducing the positive impact or increasing the negative impact on one of them[15].

Moreover, Karaca (2021), Ohnesorge (2020), and Wilholt (2009) claim that scientists must construct these models based on statistical evidence or data, as they cannot know in advance which sub-values compose a value. In fact, they need these data (inductive risk data) for mapping the correct relationship between the distinct impacts on these distinct values. This shows that scientists cannot know without statistical evidence which sub-values compose a value[16].

The need for inductive risk data that arises in the construction of inductive risk management models poses the following question: how scientists must manage this *data acquisition*? Wilholt (2009) indicates that, due to *preference bias*, scientists might *ignore* bodies of inductive risk data that are crucial for understanding the relationship between the impacts on multiple human values

[15] Consider the case of the production of a vaccine for a pandemic. The value that scientists seek to protect is health. This value is composed of two sub-values: safety and efficiency. A safe vaccine does not present toxic characteristics. Consequently, before distributing this vaccine among the persons that might be infected by the pathological agent, scientists must perform a toxicological test. If they situate the statistical threshold of this test at a higher point, the probability of the vaccine of being toxic will be higher and consumers will not be using a safe product. This event will impact their health negatively. Correspondingly, if they situate this threshold at a lower point, the probability of the toxicity of finding the vaccine toxic will be high. This will delay the distribution of the vaccine, as pharmaceutical producers will have to redesign various of its components and production methods for reducing this toxicity. This might cause more deaths by the disease and health will also be negatively affected. Thus, scientists must model the relationship between these two sub-values to find the optimal way to protect the value of health. In this case, this optimal solution would be situating the threshold neither high nor low.

[16] Wilholt (2009:93) mentions the case of the testing of the health hazards of vinyl chloride. He shows that in a first moment scientists believed that the health hazards of vinyl chloride were associated solely to toxicity and liver cancer. However, subsequent scientific studies demonstrated that these hazards were also related to other types of cancer, such as brain cancer. In other words, the pursuit of the value of health depended not just on the pursuit of the sub-values of liver health and zero levels of toxicity, but also on the sub-value of brain health.

that a scientific decision might have. Preference bias is defined by Wilholt as the fact that a specific risk management model is more beneficial for a scientist or a group of scientists[17]. For this reason, Karaca (2021), Ohnesorge (2020), and Wilholt (2009) consider that it is fundamental to regulate the *behaviour* of scientists regarding the acquisition of inductive risk data.

Wilholt (2009) proposes a first *methodological paradigm* for regulating this behaviour. It is called *methodological conventionalism*[18]. According to this philosophical perspective, all the stakeholders that represent a human value that might be impacted by a scientific decision-making process should negotiate the rules that guide the establishment of statistical thresholds in this process. In this way, these stakeholders guarantee that scientists will not prefer an inductive risk management model over another and, consequently, that they will seek all the inductive risk data necessary for understanding how all these multiple values are related. Additionally, these rules must be integrated in a *convention* that dictates how scientists should perform a decision-making process. Consequently, this convention determines which sources of data scientists should consider and which they are allowed to ignore for constructing inductive risk management models free from preference bias.

Ohnesorge (2020) finds that following methodological conventionalism is in some cases problematic. He states that this perspective does not consider the fact that conventions might be *flawed* and that, thereby, they might fail in regulating biased scientific behaviour. For this reason, conventions cannot always be used for regulating the behaviour of scientists regarding inductive risk data acquisition[19].

---

[17] For example, toxicologists working in a pharmaceutical company that is producing a vaccine for a pandemic are benefited if the vaccine is distributed earlier among the persons at risk of contagion, as this will show the efficiency of these scientists and reduce the costs of the project. Therefore, they will claim that redundant toxicological tests do not warrant the safety of the vaccine and they will ignore the data that show that redundant testing is a sub-value of safety. This shows that they will prefer a model that does not consider redundancy as a sub-value that must be preserved and might neglect the acquisition of data that corroborate that this sub-value must be protected if safety is also to be protected.

[18] The debate between methodological conventionalism and permissive empiricism can be traced back to the discussion about *measurement in science* (Ohnesorge, 2020; Tal, 2020; Wilholt, 2009). This discussion is about the role of conventions in measuring and testing processes. On the one hand, conventionalism claims that measurement is possible only if scientists use measurement conventions. Instead, operationalism considers that how scientists perform measurement operations is what legitimizes these operations.

[19] Ohnesorge (2020) and Jacob Stegenga (2017) study the case of Randomised Control Trials (RCTs). These trials are performed by scientists working in pharmaceutical companies with the aim of testing how successful a drug is. As this successfulness is determined by how much benefits this drug brings and, at the same time, how harmless it is, scientists must balance the protection of the sub-value of being beneficial and the protection of the sub-value of being harmless for preserving the value of being successful. For this reason, they perform the following procedure: in the first phase of the trial, they give all the drugs under test to the human test subjects. The drugs that produced harmful effects in them are discarded. Instead, the drugs that were harmless are used in the second phase again. In this second phase,

Ohnesorge (2020) claims that the motive that leads the stakeholders whose values might be impacted by a scientific decision-making process to regulate the inductive risk data acquisition of this process by means of a convention is *coordinative*. This means that a convention seeks to map all the values that might be impacted by this process and, consequently, that scientists, during this process, do not need to assess whether this process is preserving these values in an *optimal* way. In other words, the convention makes possible that stakeholders focus on formulating the *value judgements* that decide how a scientific decision-making process might impact human values and scientists focus on gathering the data that supports these value judgements. However, as Ohnesorge and Stegenga (2017:134-150) show with the case of RCTs, stakeholders might not know which inductive risk data is necessary for constructing the inductive risk management models that adequately preserve their human values. The reason for this is that this information can only be obtained by applying the convention. Therefore, this lack of inductive risk data might lead them to the formulation of flawed or ineffective conventions[20]. For this reason, Ohnesorge agrees with methodological conventionalism that scientists should acquire inductive risk data according to how the convention dictates. However, he states that scientists, in the cases where there are no data

scientists give the harmless drugs to the test subjects with the goal of evaluating whether the drugs produced benefits in them. In the third phase, scientists give the harmless drugs to a different group of human test subjects. This increases the probability of finding benefits. Ohnesorge and Stegenga indicate that the results of the first phase are not published because this would destroy the equilibrium between finding benefits and finding harms. In fact, as harmful molecules are product of mistaken pharmaceutical fabrication procedures, if other producers know these results, they would be discouraged of following these procedures. However, these procedures also produced the harmless and beneficial drugs tested in the second and third phase. Thus, if these other producers do not follow them, they would reduce the probability of discovering and producing harmless and, perhaps, beneficial drugs. For this reason, it is preferable to publish just the outcomes of the second and third phase, where the balance between benefits and harms is achieved. Therefore, in RCTs, the convention that guides scientific decision-making states that scientists should not publish the results of the first phase and, consequently, that these results are not relevant in the construction of inductive risk management models.

However, Ohnesorge (2020) and Stegenga (2017:134-150) show that maintaining the outcomes of the first phase out of the public light can be problematic. The reason for this is that these results do not show solely which drugs are harmful but also which groups of molecules and groups of drugs might be harmful. It is necessary to indicate that the harmful drugs are composed of molecules. Thus, their harmfulness shows the potential harmfulness of drugs produced with similar or equal molecules. Consequently, every first phase of a RCT can gather new evidence of this group of molecules and this evidence increases progressively. Nonetheless, as this evidence is kept in secret, it cannot be used for producing new drugs. Therefore, the only evidence that new producers can gather is that produced in the first phase of their trials and that gathered in reports of the second and third phases of other RCTs. This could destroy the equilibrium or balance between finding benefits and harms, as, with this evidence that is inclined to show more benefits than harms, the probability of finding benefits will increase, but the probability of finding harms will remain the same. This shows that this convention leads scientists to ignore crucial data for understanding how to balance benefits and harms, and, hence, preserving successfulness. For this reason, it is flawed and cannot regulate adequately inductive risk data acquisition and avoid preference biases.

[20] In the case of RCTs, the performance of these trials multiple times will accumulate enough evidence to assess whether the design of these trials is the optimal for balancing benefits and harms or not.

regarding which inductive risk data are necessary for constructing adequate risk management models and, consequently, formulating effective conventions (from now on I will call them *convention performance data*), such as the RCTs, should also search and acquire these data that test the effectivity of these conventions. As the acquisition of this second type of data cannot be regulated by these conventions, Ohnesorge shows that the inductive risk data acquisition process, in the cases where there are no convention performance data of these conventions, should not be regulated solely by these conventions. Rather, in these cases, this process should also be regulated by what the application of these conventions tells scientists about the effectivity of them.

This philosopher (Ohnesorge, 2020) proposes a second methodological paradigm: *permissive empiricism.* This perspective establishes that the behaviour of scientists regarding the acquisition of inductive risk data should follow two steps. First, scientists should follow the convention that regulates this data acquisition and consider the data that enable them to model the inductive risk of a scientific decision-making process according to the stakeholder negotiation that formulated this convention. Second, scientists should evaluate the application of this convention and search for the convention performance data that prove that it is flawed. If they find these data, they should give it to the stakeholders and these stakeholders should modify this convention.

The distinction between these two methodological paradigms raises the question about how to identify those scientific contexts where stakeholders do not now the effectivity of their conventions, i.e., those contexts where there are no convention performance data (from now on I will call them *poor-data contexts*), and those where there are these data (*rich-data contexts*). As the management of inductive risk is a mathematical modelling process that requires the regulation of inductive risk data acquisition through effective conventions, this practice requires the answering of the previous question. Additionally, as I have mentioned in section 3-A, the management of inductive risk must be done at the planning, internal, and external stages of the scientific process. Thus, this question must be answered at these three stages.

B) Methodological paradigms in the predictive maintenance of railways

PdM of railway systems must be *data driven* (Davari et. al., 2021; Xie et. al., 2020; Le Nguyen et. al., 2020; Chenariyan Nakhaee et. al., 2019). On the one hand, these systems operate in varying

environmental conditions and are exposed to continuous use. Therefore, it is necessary to monitor constantly the working conditions and status of the components of the system. On the other hand, the decisions when and how to maintain a component must be fully accurate, since a failure in the system might cause a catastrophic damage, that affects the life and safety of the users of the system and cause deep economic losses as well (Xie et. al., 2020). For the previous reasons, the monitoring of railway systems must progressively try to include high number of parameters in the monitoring processes. This is achieved by maintainers by increasing the methods to inspect railway systems. Multiple sources of data are used in PdM. Among them, there are sensors localized in railways and in inspection and service trains, testing methods such as ultrasonic testing, patrols of human inspectors, and monitoring instruments such as cameras and detectors. Additionally, maintainers must know how to articulate all these data sources to produce reliable monitoring results. Therefore, maintainers use *statistical models* that are aimed to infer the true relationships between all these parameters or variables, and *machine learning* (ML) models that, through regression and classification processes, extract patterns from the complex datasets produced by all these multiple monitoring techniques.

Each of these models, statistical and ML, adopt one of the two methodological paradigms explained in the previous section and, therefore, both are models used either at poor-data or rich-data contexts. In the use of statistical or stochastic models, maintainers collect data that support a PdM decision, i.e., a statement about the PdM process[21] (Xie et. al., 2020; Le Nguyen et. al., 2020;17-18). Therefore, they utilise a function that models how the changes in a variable or parameter change this statement. Thus, if the function surpasses certain threshold, it indicates that the statement that can be done by maintainers should be different[22]. This statement assuredly depends on other variables. However, as these other variables might make more complex the understanding of the relationship between the statement that can be made by maintainers and the

[21] A maintenance decision (Serradilla, et. al., 2021) is a statement about one of the three steps of the PdM process: anomaly detection, diagnosis, and prognosis. A statement about the first step would be "this rail has X abnormal behaviour". A statement about the second step would be "X is caused by the cracks that the rail has". A statement about the third step would be "this type of cracks gives the rail α days of useful life".

[22] For example, a function that models how the number of cracks of a rail degrade this component until its point of failure must represent how changes in the number of cracks increase the probability of failure of the component. Also, the function must express at which point the component enters a critical point where it can abruptly fail. Moreover, the function must represent how the time is related to the appearance of more cracks. Consequently, this function is calculating the remaining useful life (RUL) of the component based on the relation of the parameters 'time' and 'number of cracks'.

variable selected, they are excluded in the calculation[23]. Thus, stochastic modelling tries to understand the relationship between maintenance decisions and risky situations (the failure of the component) based on the selection of specific bodies of data or variables. Therefore, this type of modelling must follow conventions, i.e., predetermined assumptions about how a scientific decision impacts a human value and, hence, is risky, that regulate which data maintainers must use for designing the inductive risk management models of PdM processes. For this reason, statistical modelling in PdM adopts methodological conventionalism and is applied at poor-data contexts.

By contrast, ML modelling adopts permissive empiricism and is applied at rich-data contexts. These models have the goal of modelling the relationships between an input dataset related to a railway system component and an output dataset related to a risky situation (Xie et. al., 2020; Le Nguyen, et. al., 2020:17-18). Thus, they seek to answer the question "given these inputs, which outputs will be produced?". Additionally, these modelling practices do not use necessarily conventions that establish which variables and complementary bodies of data must be used by maintainers in the modelling of these relationships. Instead, these models examine multiple functions that map in different ways these relationships and select that function that uses more data for mapping these relationships[24]. Therefore, in the sense that ML models aim to incorporate always more parameters, they aim to use each time more data or empirical evidence to adequately understand the relationship between a group of inputs and a group of outputs. For this reason, ML modelling adopts permissive empiricism.

Until this point of the thesis, I have responded the first sub-question. In the first place, I have shown that the inductive risk that arises in the PdM of railway systems is an inductive risk that (1) must be managed through the identification of the two outcomes of scientific decision-

[23] In the previous example, the degradation of the rail assuredly depends on other variables such as how many times a train uses this rail track, whether the rail is in a flat geography or not, the environmental conditions which the rail is exposed to, among others. However, as these other variables might make more complex the understanding of the relationship between time and cracks, they are excluded in this calculation.

[24] For example, in a supervised model, modelers give to the algorithm that performs the ML modelling a set of examples of the input-output relationship. Suppose that this model maps the relationship between number of cracks in a rail and the need for maintaining the rail in the following month. Modelers can give the algorithm certain images of both rails with cracks that needed to be maintained the next month and rails that did not. Based on these images the algorithm produces a function that maps the relationship. This function states "if the image shows more than 10 cracks, the rail must be maintained next month". However, suppose that the modelers have an image of a rail that has 10 cracks but did not need maintenance in the following month. The reason for this is that the size of the cracks was exceedingly small. Thus, if the model integrates this image, then it would use additional data, i.e., the sizes of the cracks. The algorithm will produce the following function: "if the image shows more than 10 cracks that measure more than 10 cm, the rail must be maintained next month". This function is more accurate, since it can discriminate between small and big cracks and how this size affects the need for maintenance in the next month.

making processes that might impact negatively human values (false negatives and false positives). 2) As the PdM practices of railway systems have the three stages of scientific processes, this risk must be managed at each of them. 3) This management consists in responding whether the context where each of the stages of this scientific decision-making process is being performed is a rich-data or a poor-data context. In the second place, I have shown why the inductive risk of the PdM of railway systems must be managed globally and through the identification of poor-data and rich-data contexts. First, the inductive risk of the PdM of railway systems must be managed because the principal goal of this PdM is to decide whether the statistical evidence gathered from a component might be used for claiming that this component should be maintained and, in this way, mitigating a negative impact of the system on human values, such as safety, efficiency, and environmental responsibility. Second, as the principal stages of the maintenance process are three, anomaly detection, diagnosis, and prognosis, these stages can be interpreted as the three stages of a scientific decision-making process (planning, internal, and external). Third, the management of the inductive risk of this PdM consists in the distinction between rich-data and poor-data contexts, since this maintenance practices manage the impact on human values based on data modelling processes. In the next sections I will respond the second and third sub-questions.

5. Deep learning and inductive risk

A) Inductive-risk-balanced and inductive-risk-imbalanced contexts

Karaca (2021) and Biddle (2020) have used the philosophy of the societal applications of *deep learning* (DL) models[25] for examining the relationship between inductive risk and the availability of data in scientific processes. These two philosophers indicate that these models are employed in professional or societal applications where inductive risk must be managed. They also highlight the fact that regarding the management of this type of risk these applications can be divided between those where this type of risk is imbalanced and those where is balanced. Thus, they divide between *inductive-risk-balanced* and *inductive-risk-imbalanced contexts.* Inductive-risk-

[25] The discussion Karaca (2021) and Biddle (2020) do is about the societal applications of Machine Learning (ML). However, these philosophers state that DL is the most advanced form of ML. Therefore, for clarity purposes, I state that the discussion of Karaca and Biddle is a discussion about Deep Learning.

imbalanced contexts are those applications where the impact on human values of one of the false results (also called *cost*), either a false positive or false negative, is higher than the impact of the other error. Inductive-risk-balanced contexts are those applications where both errors have the same cost. For this reason, these authors claim that the aim of using DL models is, in the first place, to distinguish between these two types of contexts. Consequently, in the second place, at contexts where error costs are imbalanced, the aim of DL modelling is to balance them through the design of these models in a way that they tend to decrease the probability of the errors with higher cost. Instead, at contexts where error costs are balanced, the aim of DL modelling is to keep the probability of both types of error at the same level through the design of these models in a way that warrant that both types have the same probability of occurring. This shows that the main goal of the designers of DL models used in societal application where inductive risk must be managed is to differentiate between *cost contexts*, i.e., between inductive-risk-balanced and inductive-risk-imbalanced contexts[26].

In the previous section, I have indicated that the management of inductive risk entails distinguishing between poor-data and rich-data contexts. Thus, in scientific decision-making processes that use DL models where this type of management is required, the use of these models must be related with this distinction. I have shown that the use of DL models in scientific decision-making processes where inductive risk must be managed has the main goal of dividing these processes into inductive-risk-balanced and inductive-risk-balanced contexts. For this reason, understanding the management of inductive risk in the scientific decision-making processes that use DL models requires studying the relationship between distinguishing poor-data from rich-data contexts and distinguishing inductive-risk-balanced from inductive-risk-imbalanced contexts.

In section 3-A, I have indicated that the management of inductive risk must be done at the three stages of scientific processes, planning, internal, and external. Karaca (2021) and Biddle

[26] For instance, in the use of a DL model for oncological diagnosis a false negative has a greater impact than a false positive (Karaca, 2021). Diagnosing a patient as healthy when she has cancer (false negative) will delay the treatment of this patient and might cause her death. Instead, diagnosing a patient as ill when she is healthy (false positive) might entail solely that this patient does additional tests to confirm her truthful health condition. Clearly, for the clinic where this model is implemented the cost of producing false negatives is much higher than the cost of producing false positives. For example, if the patient dies of untreated cancer, it will have to give her relatives a huge financial compensation. Instead, performing additional tests will signify a much lower cost. In this case, DL models are used for balancing or compensating such costs. Therefore, scientists should design a model that tend to produce less false negatives than false positives by lowering the threshold of statistical evidence required for diagnosing a patient as having cancer. Instead, at a context of application where false positives and false negatives have the same costs, scientists should design a model that has no tendency to produce one type of error more than the other; the probability of errors should be the same.

(2020) reveal that the societal applications of DL can be understood as three-stages scientific processes, that have a planning, an internal, and an external stage. The societal applications of DL, as other DL modelling activities, have the main goal of analysing complex datasets and discovering patterns in these datasets. At the *training* stage, the designers of these applications introduce in these models an input dataset and an output dataset (Karaca, 2021:6-9). The model must map or calculate the relationships between these inputs and these outputs. Therefore, these designers must find a function that adequately represents these relationships between these two datasets. This is the pattern that the models aim to discover: it states, given a specific input, which output is produced. Additionally, the designers must introduce in the model a set of possible functions that map this relationship. The model must select that function that represents this relationship in the most accurate way. Therefore, as the designers must decide which functions can be used by the model for mapping the relationship, the choices they do regarding these functions determine the outcome of the model. Consequently, the training stage is the *internal stage*.

The second main goal of DL models is to *generalise* this function to other datasets distinct from those used in the training stage (Karaca, 2021:6-9). Thus, during the *testing* stage, the designers introduce additional datasets and observe whether the function found at the training stage maps adequately the input-output relationship of the new datasets. As the function is a hypothesis that specifies how the data of an input dataset is related to an output one, in the testing stage this hypothesis is tested. For this reason, the testing stage is the *external stage*.

The third main goal of DL models is to *optimise* the generalisation function (Karaca, 2021:6-9). As the exact input-output relationship cannot be found by the model because this would require the analysis of all the possible datasets that map this relationship, every function discovered by the model has an error margin. Additionally, as designers, at the testing stage, use the function discovered by the model to map new datasets, during this stage, the model can produce errors if this function does not adequately map the input-output relationships of these new datasets. Therefore, at the testing stage the model has also an error margin. If these two error margins are very distinct, it means that, during the training, the designers are introducing possible functions that do not consider all the parameters necessary for estimating the input-output relationship. Contrarily, if these two error margins are remarkably similar, it means that the model is using excessive parameters for finding the relationship. For this reason, the outcomes of the testing stage constrain the methodological decisions of the training stage; designers must find an equilibrium

between using insufficient parameters and using excessive parameters. The finding of this equilibrium is done at the validation stage. As the planning stage in science is the stage where scientists consider how outcomes constrain methodological decisions, the validation stage is the *planning stage*.

Karaca (2021) states that at each stage of the use of DL models, designers have two methods for performing the stage. At the training stage, scientists can use either equal costs or different costs for determining the possible functions they will introduce in the model (Karaca, 2021:11-12). At the testing stage, scientists can use either the ROC or the F_1 score metrics for evaluating the generalisation procedure of the model (Karaca, 2021:14-17). At the validation stage, scientists can use either a rule-based approach or a validation-set approach for finding the equilibrium between using insufficient and excessive parameters (Karaca, 2021:12-14). Since the contexts where global inductive risk must be managed are two (inductive-risk-balanced and inductive-risk-imbalanced), each of these methods corresponds to each of these two contexts. Therefore, the use of DL models in scientific decision-making processes where inductive risk must be managed consists in deciding which methods should be employed at each of the stages of these scientific decision-making processes.

Eric Winsberg employs the conceptualisation of *design* made by Wimsatt, Tebaldi, and Knutti (2012:118;128) for understanding the relationship between inductive risk management and mathematical modelling. This conceptualisation claims that design is the process of combining multiple *modelling methodologies*, i.e., distributing each of them at the distinct stages of the scientific decision-making processes where inductive risk must be managed, and using this process for distinguishing between poor-data contexts from rich-data ones at each of these three stages of these processes. DL modelling used in these processes consists in employing one of two methodologies, the one used at inductive-risk-balanced contexts and the one used at inductive-risk-imbalanced ones, at each of the stages of these processes for managing inductive risk. Therefore, scientists can utilise a design approach and, thereby, combine these two methodologies for distinguishing poor-data from rich-data contexts at each stage of these processes. Therefore, the study of the management of the inductive risk that arises in the societal applications of DL modelling consists in analysing how scientists and other stakeholders involved in these scientific decision-making processes approach this management as a design practice. Consequently, this study consists in analysing how these scientists and stakeholders employ the two abovementioned

modelling methodologies of DL modelling at each of the stages of the scientific process for distinguishing between rich-data and poor-data contexts.

B) Deep learning in the predictive maintenance of railway systems

The railway industry introduced the Internet of Things (IoT) framework (Davari et. al., 2021; Le Nguyen et. al., 2020; Xie et. al., 2020; Chenariyan Nakhaee et. al., 2019) for *monitoring* railway systems in an *accurate* and *reliable* way. The reason for this is that this information and communication technology framework indicates that it is possible to model engineering systems, such as railways, in *real-time*, i.e., in a way where every physical change in the system is reflected in the model. For example, if a rail is scratched, the model represents this physical alteration of this component of the railway system. This can be done by constructing and implementing a wireless network of varied sensing, data-processing, and data-visualisation devices, i.e., of varied *data management* devices, that acquires, integrates, and represents in an instantaneous way all the data that describes the physical status of the system in a single model (for a visual representation of the real-time railway system PdM model see figure 1; for a visual representation of the architecture of a data management network that implements the IoT framework for PdM see figure 2). Therefore, the IoT framework enabled maintainers of railway systems to monitor these systems. This is to say that these maintainers could model the *physical behaviour* of the railway systems that were constructed and regulated under this framework.

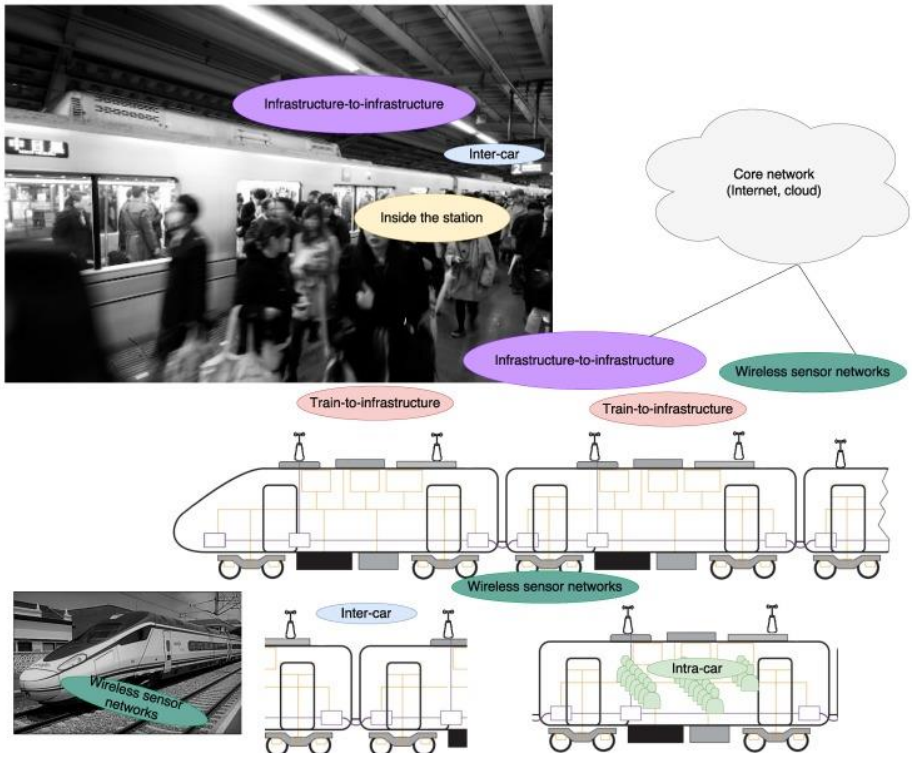

Figure 1: In this image it is possible to see how sensors embedded in the multiple components of the railway system share continuously to a unified network the data acquired from these components (Fraga-Lamas, Fernández-Caramés and Castedo, 2017:4).

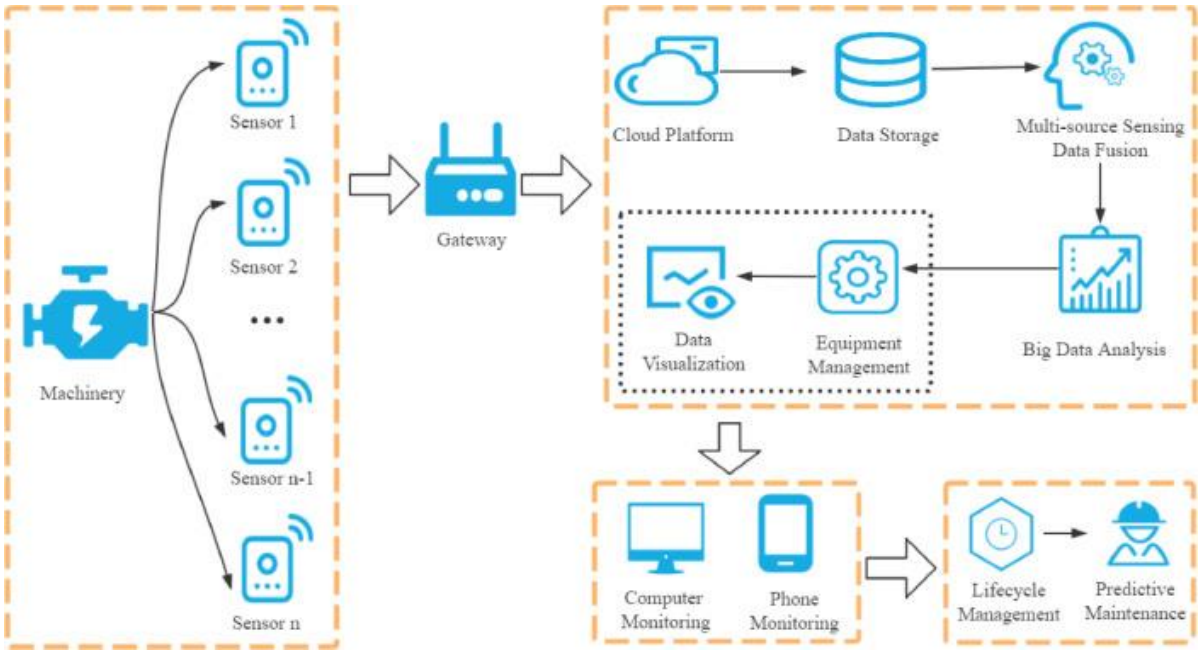

Figure 2: This image shows how the data flow from the phase of acquisition into the phase of visualization in the data networks used in systems that perform PdM, and which devices are related in this data modelling process (Huang, Liu, and Tao, 2019:102).

The use of the IoT framework in the PdM of railway systems meets the first fundamental characteristic of the DL societal applications where inductive risk must be managed. This characteristic is that the management of inductive risk entails the distinction between the inductive-risk-balanced and inductive-risk-imbalanced contexts of these applications. The railway industry uses this framework for differentiating, based on the monitoring of the physical elements of the system, the inductive-risk-balanced contexts from the inductive-risk-imbalanced ones in the PdM of railway systems and, consequently, balancing the costs of errors by managing the probability of them. In fact, the railway systems constructed and regulated under this framework reflect in the models of these systems how the physical changes in them affect their non-physical components, among them the human values that might be impacted by the PdM practices performed in them (safety, cost efficiency, environmental responsibility, usability, among others). For instance, they can represent how the scratches in a rail track in combination with its continuous use can increase the probability of a derailment and, hence, of threatening the lives of train passengers. Therefore, this framework enables maintainers to know how their scientific decision-making processes and their results which consists in physical alterations of the system impact these human values. In the previous case, an outcome can be avoiding maintenance of the scratched rail in a specific lapse of time. The important element to see is that the IoT framework shows that the costs of errors in relation to human values and the methods for balancing them depend on the monitoring of the physical components of the railway systems.

A second feature of the implementation of the IoT framework in the PdM of railways is that it created the need for *automatic pattern discovery* in these practices (Davari et. al., 2021; Le Nguyen et. al., 2020; Xie et. al., 2020; Chenariyan Nakhaee et. al., 2019). Previous to the implementation of this framework in these practices, the modelling of these systems required pre-processed datasets. This means that maintainers decided and selected one by one, in a manual way, the parameters necessary for modelling the relationship between physical changes in the systems and maintenance actions and the impacts on human values. Therefore, the inferential relationship between a maintenance action and a physical change and an impact depended on known features or parameters. After the parameters were selected, maintainers, through *inspection* activities, extracted or acquired the data that described the physical status of the system according to these parameters. However, with the introduction of data management devices (sensing, data-processing, and data-visualisation devices), the amount and combinations of data increased continuously and

became so vast that maintainers could no longer manually select these parameters. During inspection, DL models automatically select these parameters in this ever-changing data stream and analyse them for discovering patterns that describe the inferential relationships between the physical status of the system, maintenance actions, and the impacts on human values. Therefore, the IoT framework and the new data-management technologies transformed inspection into an automatic process. In this new process the data necessary for managing inductive risk is not acquired according to pre-selected parameters, but on models that, through the combination of multiple groups of parameters, found the best way to represent the impacts of the railway system on human values.

This use of automatic pattern discovery shows that railway systems constructed and regulated under the IoT framework meet the second fundamental characteristic of DL societal applications where inductive risk must be managed. As indicated in the previous section, this characteristic is that the balancing of error costs must be employed by DL modellers to decide whether the context where inductive risk is being managed and where the DL societal application is being deployed is a rich-data or a poor-data context. As the IoT framework transformed the management of inductive risk into a process that consists in relating the physical changes in the system with the impacts on values, the data necessary for managing inductive risk are data that describe the physical characteristics of the system. For instance, a dataset used for predicting the failure of a rail track and its relation to the scratches of the rail contain data about the three-dimensional configuration of the scratch. The dataset is made from numerical data about the height, width, length, size, and other variables necessary to represent this configuration (Santur, Kakaröse, and Akin, 2017). I have highlighted that DL models combine different parameters and, based on the data acquired during the inspection of the component of the system, discover patterns of failure and their relationships to the physical characteristics of this component described by these data. Therefore, the DL models do not rely on predetermined parameters but on the combinations between the parameters found in the data for finding these patterns. In this way, during the use of these models, maintainers employ the data gathered by the inspection activities, i.e., empirical data, for managing inductive risk and transforming the data contexts where the maintenance practices are performed from poor-data into rich-data.

The DL models used in the PdM of railway systems use multiple *architectures* or modelling methodologies for automatically discovering patterns (Davari et. al., 2021; Le Nguyen et. al., 2020; Xie et. al., 2020; Chenariyan Nakhaee et. al., 2019). Also, these architectures are classified according to the type of dataset that they use for automatically discovering patterns. These datasets are classified into two types: 1) the datasets composed of images and 2) the datasets composed of single measurements. The main difference between these two types is that, while in the use of the first type the DL model needs to analyse the representation done by the image of the physical status of the component of the system for discovering the pattern, in the use of the second type it needs to analyse a specific measure captured by the inspection device. As an image represent the multiple measurements applied to a component (e.g., the colour, the area, and the shape of the component), the main difference between these two types of architectures is that, whereas one uses a single measurement for constructing a dataset, the second uses multiple measurements. For this reason, DL architectures are classified in those that use single measurements for finding classification patterns (from now on I will call them *separated architectures*) and those that use multiple measurements (from now on I will call them *integrated architectures).* Additionally, the use of DL models for PdM of railway systems consists in employing these two types of DL architectures of automatic pattern discovery for detecting anomalous behaviour or conditions in the components of the system, classifying these anomalies into distinct types of faults, and prognosing the RUL of these components[27].

This classification of architectures meets the third characteristic of DL societal applications where inductive risk must be managed. This third characteristic is that in these applications modellers combine two modelling methodologies, the ones used at inductive-risk-balanced and the ones used at inductive-risk-imbalanced contexts, and use this combination for distinguishing rich-data from poor-data contexts at the three stages of these scientific processes. As mentioned earlier, automatic pattern discovery transforms poor-data contexts into rich-data ones and, thereby,

[27] Both technical writers (Chenariyan Nakhaee et. al., 2021; Davari, et. al. 2021; Xie et. al., 2020) and philosophers (Karaca, 2021:19-23; Erasmus, Brunet, and Fischer, 2020) acknowledge that DL models have multiple architectures, such as the convolutional neural network (CNN), the recurrent neural network (RNN), and the generative adversarial network. The difference between these architectures lies on how they utilise and combine multiple mathematical functions for modelling the relationship between the input dataset and the output dataset at the different points or nodes where this relationship is modelled. In the case of the use of DL for the inspection activities of the PdM of railway systems, the architecture of DL models depends on the articulation of the multiple elements of these activities: which data are used, how these data are acquired, and which is the influence of these data on the decisions regarding maintenance. Thus, these elements classify these architectures into two types, integrated and separated.

manages inductive risk by combining the multiple parameters that describe the data gathered at the inspection activities. The construction of datasets composed of multiple measurements decreases the probability of one type of inductive error. The reason for this is that using two distinct measurements for finding the same pattern decreases the chance that the model fails in finding this pattern. For instance, if a model is used for detecting a faulty rail track and it uses both measurements about the structural properties of the rail and images of the rail, it has greater chances of truly detecting this failure than a model that uses solely a single measurement. Since the management of inductive risk depends on the finding of patterns, the use of integrated measurements for automatic pattern discovery depends on the costs of the failures to be discovered. Therefore, in the cases where these costs are the same, maintainers do not need to use a method that decreases the probability of producing one type of errors. Thus, in these cases maintainers can use DL separated architectures. By contrast, in the cases of costs imbalances, maintainers must use integrated architectures and reduce the probability of the errors with higher costs. In this way, maintainers use a design approach and combine the multiple DL architectures in order to manage error costs and, hence, inductive risk at the three stages of the scientific process where inductive risk is to be managed.

Chapter 4: Inductive risk, Deep Learning and Testing

In this chapter I analyse the external or testing stage of the use of DL. In the first place, I observe how the use of testing methodologies is decided by the cost contexts where DL is used. In the second place, I assess how this distinction between cost contexts can justify the methodological testing decisions taken by the engineers who proposed the two solutions selected.

1. Value contexts and testing of deep learning models

DL models are tested according to how well they classify a new set of data based on a general pattern found in the training data (Karaca, 2021:6-9;14-17). The solutions to this generalisation problem are multiple, since they depend on which parameters are used by the algorithm in order to classify datasets. Therefore, the methodological decisions taken by the designers of the DL model are the aspects of these models that are tested. According to Karaca (2021:14-17), there are two ways for testing the generalisation process done by DL models. The first one is applying the metric known as receiver operating characteristic (ROC) that subtracts the ratio of all false positives divided by all negatives *FP/N* (false positive rate or *FPR*) with the ratio of all true positives divided by all positives *TP/P* (true positive rate or FPR; also called sensitivity). The formula is: *TPR-FPR*. The second applies the F_1 score metric. This application consists in combining the result of this subtraction with the precision ratio, which is the division of true positives by all the instances classified as positive (*TP/TP+FP*). The formula this second metric use for combining these two quantities is: *2 * (Precision*Sensitivity)/(Precision+Sensitivity)*. Thus, the F_1 score is formulated in the following way: $F_1 = 2 * (Precision*Sensitivity)/(Precision+Sensitivity)$.

Karaca states that the F_1 score metric should be used when the multiple model choices use highly imbalanced classes (2021:14-17). This means that most of the features of the dataset selected by the model for performing the generalisation procedure are not used by the other models. In this way, the greater the difference between the features selected by each model, the greater the use of

imbalanced classes by these models. Karaca also states that designers select these features based on the values or interests these scientists aim to defend with the construction of these DL models[28].

These two claims done by Karaca show that the methodological choices for testing DL models depend on the relationship between the different human values designers seek to obtain with the implementation of these models and the costs or impacts on these values of the errors that can be produced by these models. If the errors have balanced costs, then they should use the ROC metric. Contrarily, if these errors are imbalanced, they should use the F_1 score. Therefore, I observe that, while the ROC is used in inductive-risk-balanced contexts, the F_1 is used in inductive-risk-imbalanced ones.

In the following section, I analyse how these contexts are distinguished in the testing procedures of DL models used in the PdM of railway systems. For doing this, I analyse which criterion maintainers use for applying the ROC or the F_1 score metric in both selected cases. As there are other metrics besides ROC and the F_1 score (Hossin and Sulaiman, 2015), but that do not have relevance in the present philosophical discussion about cost contexts, I analyse the criterion maintainers use in these cases for distinguishing between inductive-risk-balanced and inductive-risk-imbalanced contexts even if they do not mention these two metrics explicitly. The association of these two metrics with the two types of cost contexts (inductive-risk-balanced and inductive-risk-imbalanced) is done for showing that maintainers follow testing choices that can be classified into two groups and that these groups are defined by the type of cost contexts where the DL model is being implemented. For this reason, the analysis in the following section is centred on how maintainers make testing choices based on how they differentiate between inductive-risk-balanced and inductive-risk-imbalanced contexts.

[28] For example, in the modelling of an oncologic DL model, scientists must choose a particular set of features in the visual characteristics of a tumour for determining how this algorithm will classify this tumour between the different classes of tumours. As some tumours require more urgent medical intervention than others, depending on this classification, a patient could receive treatment in less or more time. Thus, if these scientists are interested in preserving patients' safety, they must choose a set of features that make the model classify most of the tumours in those categories that require urgent intervention.

2. Testing of DL models

Case 1 - Digital Camera Inspection

As mentioned in chapter 2 (see section 4), the goal of this inspection system is to aid maintainers in their decisions about when they should perform a maintenance action on a rail track. Maintainers should replace missing fastening bolts for avoiding critical situations. Therefore, detecting missing bolts indicates them whether a bolt is missing and requires replacement. In chapter 2 I also have mentioned that manual inspection is slow and subjective. Thus, the introduction of this automatic inspection system reduces the number of falsely detected missing bolts (false positives). For this reason, the goal of the DL model used in this system is to reduce these false positives that occur in the detection process of this inspection activity that consists in examining whether bolts are correctly positioned[29].

Marino et. al. (2007:425-428) tested how well this model generalises its classification function by utilising images of both hexagonal-headed and hook-headed fastening bolts recorded by a video camera[30] . This test shows that the rate of false positives was significantly higher in the detection of hexagonal-headed occluded bolts (0.1% bolts were detected as correctly positioned) than in the detection of left hook-headed (47%) and right hook-headed (31%) bolts. The authors of the DL solution stated that this difference in rates is originated in the fact that the hexagonal-headed bolts are very similar to the stones of the ballast where the track is placed. As this factor was not considered by the authors for giving a higher cost to this false positive, i.e., classifying an occluded

[29] It is fundamental to indicate that aiming to decrease the number of false positives do not necessarily entail that this type of error has a higher cost in the DL model. The reason for this is that the model aims also to use a specific classification method and specific sets of data and of measurement procedures for understanding the relationship between these data acquisition processes and the probability of errors. Therefore, the costs of the errors depend on the relationship between an elected data acquisition process and how this election impacts the probability of each of these errors.

[30] In the case of hexagonal-headed bolts, the test consisted in examining a rail network that contained 3350 bolts (2469 visible bolts; 721 occluded bolts; 21 absent or missing bolts). In the case of the hook-headed bolts, the test also consisted in examining a rail network but, in this case, this network employed hook-headed bolts both with the hook directed towards the left and with the hook directed towards the right (the authors do not specify neither how many bolts the network contained nor how many were visible, occluded, and absent). The results of the test shows that the percentage of detection of left hook-headed and right hook-headed visible bolts was of 100% and the percentage of hexagonal-headed visible bolts was of 96%; the percentage of detection of left hook-headed bolts was 47%, of right hook-headed occluded bolts was of 31%, and the percentage of hexagonal headed bolts was of 0.1%; the percentage of detection of absent left hook-headed and right hook-headed bolts was of 100% and the percentage of hexagonal bolts was of 95%.

hexagonal-headed bolt as a missing one, and, consequently, reducing the probability of this type of errors, in this testing stage the errors were balanced[31]. For this reason, it is an inductive-risk-balanced context.

The testing in this case is based on detecting a shape at a specific set of coordinates within a region (Marino et. al., 2007). In fact, the system searches for the bolt in the x axis at the distance intervals established by the measurement of the distance between the first detection of a bolt image and the second detection. However, the model does not discriminate between the different outlines of the shapes of the hexagonal and hook bolts. The type of the architecture of this model is a multilayer perceptron neural classifier (MLPNC) which performs better than other DL modelling methodologies in classifying images based just on topological features, e.g., the fact that a shape is inside a region. Consequently, the designers of the model can determine whether this model generalises well its classification function by determining how well it can detect a shape in a specific position or set of coordinates without using the outline of the shape as a parameter for making this detection. In other words, the geometrical features of the image are not required for deciding whether in a segment of the rail track a bolt is missing or not. Thus, the assumption that guides this testing method is that the outline (a geometrical feature) of a shape is not necessary for detecting the existence of a shape in an image if this image occupies the entire inspection area. The topological features perceived in the image are sufficient for detecting this shape. This shows that the DL model uses solely one type of measurement (topological features) for managing the inductive risk of the testing stage. Therefore, at the testing stage of this DL solution, the designers of the model use a DL architecture that measures just one feature for maintaining the probabilities of errors balanced.

Marino et. al. (2007) indicate that the rate of false positives at the training stage is higher than this same rate at the testing stage, since the model is trained with images of rail tracks that use hexagonal-headed bolts and this type of bolts tends to be misclassified as stones and, hence, as missing bolts. Instead, the rate of false positives is lower at the testing stage, since the model, in addition to hexagonal-headed bolts, uses hook-headed ones who have lower probability of being

[31] It is important to mention that at this test stage the test dataset is composed of the images of the hexagonal-headed, the left hook-headed and the right hook-headed bolts. Thus, the false negatives produced in the testing of each of these subgroups of images must be summed to establish the general rate of false negatives. As the designers did not consider the fact that hexagonal-headed occluded bolts have lower chances of being detected and did not arrange the test for increasing this probability, the general rate of false negatives is not more costly for the designers than the rate of false positives. For this reason, this testing context is an inductive-risk-balanced one.

misclassified as stones. Thus, the probability of this false detection decreases. This shows that the testing method guarantees that the probability of false positives will always be lower at the testing stage than at the training stage. In other words, designers can guarantee that the probability of false positives will never be higher at the testing than at the training.

I have indicated that the testing context of this DL model is an inductive-risk-balanced one, as false positives do not have a higher cost than false negatives and, consequently, the probability of the two types of errors can be kept unaltered. Therefore, as the designers of the model can keep the probability of false positives under a certain threshold by measuring solely the topological features of bolts, they know that they are at an inductive-risk-balanced testing context because they can use a separated DL architecture for keeping the probability of this type of error under this threshold. Thus, this possibility of keeping the probability of this type of errors under a certain threshold by using a separated architecture is the feature of the system that reveals designers that they are at an inductive-risk-balanced testing context.

Case 2 - Laser Camera Inspection

Santur, Karaköse, and Akin (2017) explain that for testing how well the model generalises they use three-dimensional (3d) graphics that represent how the surface of the rails are embedded in a 3d space[32]. Thus, the testing dataset is composed of these 3d graphics. They use three-dimensional similarities between the rail surfaces data and the 3d graphics for testing how well the model generalises its classification function because they consider that the faults in the rails are changes in the standard embeddedness of the surfaces of these rails. For this reason, using 3d graphics that represented faulty and healthy rails based on the variations in the embeddedness of the surfaces of these rails is accurate for testing the classification function of the model (see figure 3).

[32] These authors also evaluate the velocity of the algorithm for detecting faults, and for this procedure they use the ROC metric. However, I consider that this test is not directed toward the model, but toward its implementation. In this second test, the classification function is not being tested but how well the classification performs when the laser moves at a certain speed (100km/h) (Santur, Kakaröse and Akin, 2017). For this reason, this second test is not discussed.

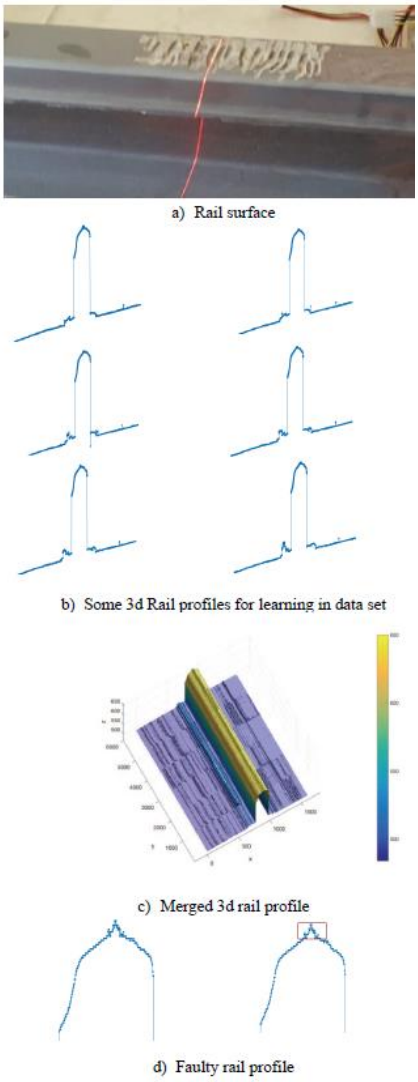

Figure 3: This testing method is based on the geometric properties of three-dimensional images. As the laser is used for determining the presence of objects at different heights and the camera is used for detecting how these objects are distributed in the plane where the rail is located, the laser camera can produce a three-dimensional image of the rail. Depending on how the surfaces that define the limits of the rail are embedded in the three-dimensional space, the model detects whether the rail presents a fault or not. The reason for this is that a fault, such as a scratch, changes how these surfaces are embedded. For instance, if two rails with the same structure, one scratched and one in good condition, are compared, they will show that the way their surfaces are embedded in the three-dimensional space is not the same in each case. The reason for this is that the surface of the scratched face of the damaged rail has a different embeddedness if compared to the surface of the same face of the normal rail. Therefore, the assumption that guides this testing method is that the examination of the three-dimensional embeddedness of the surfaces of the rails is the fundamental parameter needed for detecting a fault in these rails. Consequently, the three-dimensional properties that are not related to this embeddedness are not important for this detection. This image shows how the system constructs the three-dimensional embeddedness of the track surface by aligning the measurements of depth done by the laser at each of the segments of the track (Santur, Kakaröse, and Akin, 2017).

Santur, Kakaröse and Akin (2017) state that the testing procedure consists in classifying the rails as healthy or faulty. Since these authors does not use datasets that represent physical rails but datasets of 3d graphics, in the testing procedure they must evaluate (1) how well the model generalises its classification procedure. Also, they must evaluate (2) how well the model generalises its classification by using a specific type of measurement: the three-dimensional embeddedness of the surfaces of physical and graphical rails. In this way, the testing procedure can produce four types of results: 1) the model finds a fault in the graphical rail and this fault is present in the physical rail represented by the graphical one (true positive). 2) The model does not find any fault in the graphical rail and the physical rail represented by this graphic is also healthy (true negative). 3) The model finds a fault in the graphical rail, but this fault is not present in the physical rail (false positive). 4) The model does not find any fault in the graphical rail, but the physical rail is faulty (false negative).

This testing is an inductive-risk-imbalanced context. The reason for this is that false positives have a higher cost than false negatives. In fact, this inspection model is aimed to decrease the number of healthy rails classified as faulty. Also, it uses the measurement of three-dimensional properties for reducing these false positives. In the testing procedure, the designers do not use physical rails but graphical ones for evaluating how well this model classifies rails. Thus, they are evaluating both whether the function of the model generalises well and whether measuring three-dimensional properties is an adequate method for classification. Moreover, whereas false positives indicate that measuring three-dimensional properties is adequate, since through the analysis of these measures the model can find faults, false negatives indicate the opposite. Furthermore, both false positives and negatives indicate that the model does not generalise well. Therefore, as false negatives indicate that measuring three-dimensional properties is inadequate for finding faults and that the model does not generalise well, this type of error has a higher cost.

Since maintainers do not aim solely to test the classification model but the type of measurement it uses for making the classification (three-dimensional properties), in this case, the criterion for distinguishing inductive-risk-balanced and inductive-risk-imbalanced contexts at this testing stage is based on which are the aims of this testing. The aims in inductive-risk-imbalanced contexts are two: (1) testing the classification model by evaluating how well it generalises and (2) testing whether the measurement it uses is adequate for classification. Santur, Kakaröse, and Akin (2017) claim that the model uses a convolutional neural network (CNN) architecture at the testing

stage (see figure 4). As previously mentioned, errors have distinct costs at this testing stage, but the model uses just one type of measurement for classification. As the methodology that works for inductive-risk-imbalanced contexts is one that uses two measures, the methodology must be able to transform this measurement into two. This is achieved by the CNN, since it uses only a simple group of characteristics for classifying but reinforces this classification criterion by using complex datasets; both simplicity and complexity are features that it uses for classifying. Therefore, the CNN searches for common features in the datasets it classifies (first measurement) and evaluates the degree of complexity of these datasets (second measurement). For this reason, this DL solution shows that the possibility of applying a CNN to a testing procedure indicates the designers of the DL model for the PdM of railway systems that the testing context is inductive-risk-imbalanced.

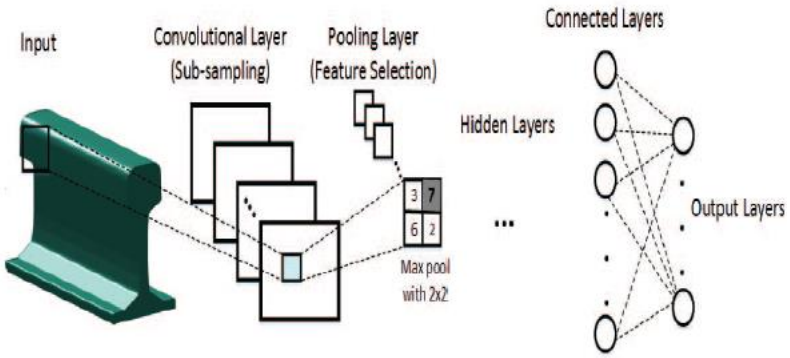

Figure 4: The fundamental characteristic of CNN architectures or methodologies is that, in a first step, they decrease the characteristics necessary for performing a classification (sub-sampling or convolution). In a second step, they use these characteristics for classifying datasets with high number of dimensions (pooling). The model performs this process until it can find functions that use a low number of characteristics for classifying complex datasets and at the same time it does not suffer great perturbations or high rate of errors (Santur, Kakaröse, and Akin, 2017).

3. Comparison and contrast

In both cases, the testing procedure uses solely part of the parameters that characterise the training and testing datasets for evaluating the generalisation ability of the model. In the first case, this parameter is how much a shape occupies a specific region. The outline of the figure is an excluded parameter. In the second case, this parameter is the three-dimensional embeddedness of the surfaces of the rail. The material characteristics of the rail are excluded parameters. In both cases, the testing data does not represent the same type of object represented in the training data. In the first case the training data represents hexagonal headed bolts, and the testing data represents hook

headed bolts. In the second case, the training data represents physical rails, and the testing data represents graphical ones. For this reason, these parameters selected in both cases are characteristics shared by both the training and the testing objects. As these parameters help maintainers to distinguish whether the testing context is an inductive-risk-balanced or an inductive-risk-imbalanced context, in both cases the criterion for this distinction is based on the evaluation of the common features shared by the testing and training datasets.

A difference between cases is that, while in the first case, the testing procedure is not evaluated, in the second case it is. Additionally, in the second case, the fact that the testing procedure is also evaluated is the criterion that allow maintainers to know they are on an inductive-risk-imbalanced context. In the first case, the fact that the testing procedure is not evaluated is not necessary for determining that maintainers are at an inductive-risk-balanced context. As far as the model does not surpass a threshold in the general rate of classification error, the maintainers can establish that the classification model generalises well. Therefore, in the first case, the criterion for establishing that they are in an inductive-risk-balanced context is that they can use a statistical threshold for determining the acceptable rates of error, without having to assess whether the objects represented in the testing data and the objects represented in the training data are sufficiently similar for legitimising the testing procedure. Instead, this legitimation is evaluated in the second case, since maintainers are evaluating whether graphical representations of rails can be used for testing models that classify physical rails.

Chapter 5: Inductive risk in Deep Learning and the planning stage of science

In this chapter I analyse the planning or validation stage of the use of DL. First, I show that each validation methodology is used at each type of cost context. Second, I reveal how the validation processes used in the two chosen solutions can be decided by distinguishing the cost context types where these solutions are deployed.

1. The validation of deep learning models and value contexts

As I mention in the third chapter, societal applications of DL, as other scientific processes, have three stages: the planning or validation stage, the internal or training stage, and the external or testing stage (Karaca, 2021). In the *training* stage, the designers of the DL model propose a classification function. For doing this, they introduce a set of *proposed functions* that approximate to the unknown *ideal function* that can map all the input-out relations or correct classifications. Also, they introduce a group of data, or a dataset, composed of inputs and outputs. The ideal function is unknown, and the proposed functions can just approximate to it because the data used for training the model is limited. Therefore, the model discovers a function based on incomplete information. Thus, the model, based on the analysis of this dataset, tries to find which function of the set of proposed functions is the one that approximates the most to the ideal function. As each of the proposed functions approximate in a specific way to the ideal function, they have various levels of approximation. Because the proposed functions can only approximate to the ideal function, designers expect that each of the functions produces classification errors. Also, as this error rate is inversely proportional to the level of approximation of the function, designers can calculate this rate based on the approximation level. Additionally, as designers know the data used by the model to select the function, they can know how much the function approximates to the ideal function when each of the inputs of these data is analysed by the model. This allows the designers to calculate the average of the approximation and, thus, of the error rate of each of the proposed functions. This average is called the *loss function* (Karaca, 2021:6-9).

At the *test* stage, designers evaluate whether the selected classification function is the function that effectively approximates the most to the ideal function. This observation consists in introducing a new dataset and performing the same selection process. The difference in the loss function of the training stage and the loss function of the test stage shows designers how well the model can use its classification system for new datasets. In other words, it shows how well the model *generalises* this classification pattern. This comparison allows designers to know whether they must change the proposed functions or not (Karaca, 2021:6-9).

A significant difference between both loss functions, the training loss function and the test loss function, indicates that the proposed functions might be ignoring key features of the dataset that correctly map the input-output relation. This problem is called *underfitting*. Correspondingly, a minor difference between the two loss functions indicates that the model is using features of the data that are not necessary for correctly mapping the input-output relation. This problem is called *overfitting*. Therefore, the main goal of designers is to propose functions that keep an equilibrium between using excessive features and not using enough features for correctly mapping the input-output relation. This process is done at the *validation* stage (Karaca, 2021:6-9).

Additionally, underfitting errors can be considered false negatives because, on the one hand, both loss functions are very distinct (Karaca, 2021:6-17). This means that the test procedure considers that the training loss function is false. On the other hand, as this training loss function was produced using insufficient data, it did not use the adequate threshold of statistical evidence. For this reason, its rule of acceptance is false. Correspondingly, overfitting errors can be considered false positives because, in the first place, both loss functions are similar. This means that the test procedure considers this loss function as true. However, as the training procedure used excessive data for producing this loss function, it did not use the adequate threshold of evidence.

As the outcomes of the testing stage decide whether the model is underfitted or overfitted and, hence, which methodological decisions should be taken by designers at the training stage for reducing these two types of testing results, these methodological decisions done at the training stage are constrained by the expected outcomes at the testing stage. In the third chapter, following Karaca (2021), Ohnesorge (2020), and Wilholt (2009), I have said that, at the planning stage of science, scientists must evaluate how the results they aim to produce with the performance of the scientific process should decide their methodological decisions at the internal stage. As taking a certain methodological decision does not necessarily assure that the outcome sought by this

decision will be obtained by scientists, these scientists must evaluate between multiple methodological decisions and select those that have higher probability of producing the result they expect to produce. This evaluation is done at the planning stage. Since, at the validation stage, the designers of DL models must choose between different methods for producing a model that is neither underfitted nor overfitted, the validation stage corresponds to the planning stage; designers must evaluate which is the methodological choice that have the highest probability of producing the perfect model, i.e., a model that, at the testing stage, will be neither underfitted nor overfitted.

Avoiding underfitting and overfitting can be done through two principal methods (Karaca, 2021: 6-9;11-12). 1) Fixing the number of dimensions used by the functions for performing the classification process. This fix is achieved by setting rules that determine the parameters all the functions can use for making the classification. 2) Estimating the test function loss by creating a validation dataset that has data of the training dataset and of the test data set. As this new set is different from the training set and from the test one, designers can predict, before testing, which will be the test loss function and how well the model will generalise[33].

The first method is used at inductive-risk-balanced contexts and the second method is used at inductive-risk-imbalanced contexts. Designers fix the number of dimensions used in the classification process with the aim of assuring that, if the test loss function is too similar or too different from the training loss function, this lack of balance in the use of dimensions is not a problem that exists in the ideal function but in the proposed functions. The ideal function ideally classifies all the data in the correct category and maps perfectly the input-output relationship between an input dataset and an output dataset. Therefore, all the proposed functions that approximate to this ideal function must imitate the input-output relationship mapped by this ideal function as much as possible. Thus, if a function uses more dimensions than the fixed, it is producing a false result. The same happens with functions that use an insufficient number of dimensions. As designers fix this number of parameters, they are establishing that underfitting and overfitting have the same error cost or weight. For this reason, false positives (using excessive

[33] Validation is a stage that not necessarily occurs always before or after the training stage. It can occur before the training stage for mitigating the risk of training a underfitted or overfitted model. However, it can occur after the training stage for understanding why the trained model is underfitted or overfitted (Wang and Zheng, 2013).

number of parameters) and false negatives (using insufficient parameters) have the same cost or weight; one of this group of errors is not prioritised over the other[34].

In contrast, if designers use a validation dataset for validating before testing how well the model generalises, they are assuming that the testing procedure can also choose the functions proposed in the training stage. Therefore, if the loss function of the validation procedure is like the training loss function, this means that overfitting actions have a bigger weight than underfitting actions. Contrarily, if the loss function of the validation procedure is significantly distinct from the training loss function, this means that underfitting actions are more problematic for finding an adequate model. For this reason, as inductive-risk-imbalanced contexts are contexts where errors or false results have different weights, the contexts where designers use validation datasets for validating models are inductive-risk-imbalanced.

In the following section I analyse how these validation contexts are distinguished in the validation procedures of the application of DL models in the predictive maintenance (PdM) of railway systems. I will examine the criteria maintainers use for applying either the first or the second method exposed above. In the same way as the earlier chapter, in the cases I study maintainers do not mention explicitly which validation methods they apply. However, it is clear how they avoid underfitting and overfitting problems. Therefore, I analyse how they avoid these problems, and the criteria for applying the methods they used for avoiding them.

2. Validating deep learning models in the predictive maintenance of Railway Systems

Case 1 - Digital Inspection Camera

Marino et. al. (2007:418-419) show that there are two types of applications of DL to visual inspection systems. On the one hand, there are DL modelling methodologies that recognise the geometrical shape of the object under inspection. For example, systems used for detecting missing fastening bolts in railways detect the geometric shape of the head of the bolt. On the other hand, there are methodologies, as the one exposed by Marino et. al. (2007), that detect whether a shape

[34] For example, in a DL model used for oncological diagnosis, designers can fix the dimensions for classifying images of tumours. They say that colour, tumour diameter, and shape are the three dimensions that should be used for classification. Thus, using less dimensions or using more will produce errors that have the same weight.

is positioned within a specified area. The main difference between these two methodologies is that, while the first needs that a human operator specifies the geometrical shape that must be searched by the DL model, the second does not need this specification[35].

Given these characteristics of these two methodologies, I can show that the first one uses a validation dataset[36], i.e., the second method of validation, for validating the classification

[35] Costa and Gonzaga (1996) claim that in automatic visual inspection there are two main types of DL architectures. The first one is called moment invariant (MI) and it learns by defining a feature in the image (e.g., a specific geometrical shape) as a reference point that must be contained by all the images that will be classified into a specific class. Thus, the DL model uses this reference point as a moment, i.e., a segment in a mathematical function, that must be possessed by all the functions that represent the images that are part of the same class. For this reason, the architecture, first, searches this moment and, second, searches for the other moments of the optimal classification function. This point of reference is defined by the designers. Consequently, the moment that mathematically represent this point must be defined by the designers each time the model is trained.

By contrast, the second type (called multilayer feedforward neural net or MFNN) do not utilize a fixed reference point for the classification. Instead, it propagates the numerical data throughout the layers of the net, from the input layer to the output layer, computing it in the hidden layer or layers according to the modelling choices done by the designers. At the output layer it compares the function it found with a set of classification examples. If the found function uses the same classification patterns as those of the examples, the training finishes. If these patterns are different, the model adjusts, in what is called a *backpropagation* action, these patterns with the aim of making them equal. Thus, it adjusts the modelling choices introduced by the designers in the hidden layer(s) (Costa and Gonzaga, 1996). The important element to highlight in this second type is that the designers do not need to fix any reference point and moment each time the model is trained. Thus, in contrast to the first type, the training of the model does not require human intervention.

Marino et. al. (2007:418) establish that their solution utilises the MFNN architecture. In fact, the DL model of this solution does not recognise the shapes of the bolts in the images by utilising a fixed point of reference or feature in the training procedure. One of the advantages of their model is that it does not require *tuning*, i.e., that a human designer define this feature each time the model is trained.

Instead, the second type of DL application in visual inspection they mention uses the MI architecture. Stella et. al. (2002) state that, besides the DL applications that do not use predefined features in their training for discovering their classification functions, there is other type of application that uses these features. Marino et. al. indicate that in this other type of application human designers must *tune* the model. This means that they must select the features used by the model for discovering its function. For this reason, there are two basic types of DL applications in visual inspection when they are classified by the DL architecture they use. The first type are those applications that use a MFNN architecture. The second are those that use a MI architecture.

[36] Marino et. al. (2007:418) refer to their previous work for distinguishing the two abovementioned types of DL models applied to visual inspection. In this previous work they claim that both models use pre-processed image features. These features (Weinmann, 2013:1-13) can be classified into four main types: 1) colour and intensity, 2) shape (e.g., perimeter, compactness, eccentricity, polygonal approximations), 3) texture (e.g., uniformity, density, roughness), 4) local features, i.e., features that appear in specific segments of the image (e.g., convex blobs, points of inflection, junctions, or intersections). The pre-processing (Marino et. al. 2007:420) consists in turning these features into numerical quantities and store them in the datasets that will be analysed by the model. As each of these features produces a specific visual signal, they can be converted into numerical quantities by using mathematical techniques that analyse the specific features of these signals or waves. Thus, the mathematical features of the wave (amplitude, frequency, and time) are used for classifying it as a specific image feature. For instance, waves that have *x* amplitude and *y* frequency at *a* time are convex blobs. Therefore, the datasets used by the DL visual inspection models are composed of these numerical quantities that represent the image features. Also, the construction of these datasets is done by signal processing mathematical techniques that transform these signals into numerical quantities that are classified according to the correspondence between a wave feature and an image feature. For instance, the datasets used by Marino et. al. (420) utilise wavelet descriptors or transforms that are mathematical functions that describe the outline of a shape. Thus, in the pre-processing, the visual inspection system searches for these functions in the images

procedure. As the solution proposed by Marino et. al. uses this first methodology, I use it as an example for showing my point. As mentioned in the footnote 35, this methodology uses a MFNN architecture. For this reason, the model produced by this methodology must use during the training process datasets that work as classification examples. Though not mentioned by Marino et. al., in the solution they propose these datasets must contain images of bolts with multiple types of head, since using images of bolts with a single type of head might lead to overfitting the model[37]. Marino et. al. (2007:425-428) mention that at the training stage they use images of hexagonal-headed bolts and at the testing stage they use images of both hexagonal-headed and hook-headed bolts. The reason for this is that they state that the model must classify bolts with multiple types of head because railways do not always use bolts with the same type of head. Therefore, a validation dataset, i.e., a dataset that contains training and testing data, is, in this case, a dataset that contains both images of hexagonal-headed bolts and of bolts with other types of heads. Since the architecture of this solution, as in other DL applications in visual inspection that use this first methodology, is a MFNN, the training of this model requires using this validation dataset; the images of hexagonal-headed bolts are the images used by the model to find the pattern and the images of bolts with multiple types of heads are used as training examples[38]. As this solution use this validation dataset, it uses the second method of validation: using datasets that contains training and testing data[39].

This solution proposed by Marino et. al. shows that in the design of DL models for the PdM of railway systems the criterion for distinguishing at the validation stage between inductive-risk-balanced and inductive-risk-imbalanced contexts can be found by designers in the possibility these designers have of using the architecture of the model for validating it. Marino et. al. state that their

---

and count how many of them exist in these images. Subsequently, it stores this quantity in the dataset under the parameter 'wavelet descriptor'.

[37] Images with bolts of just one type of head might lead the model to use the geometry of the head as the pattern for classifying a bolt as positioned. However, as there are bolts with other types of head, this classification function cannot be generalized because it uses excessive parameters (the geometrical parameters).

[38] Marino et. al. (2007) do not mention specifically whether the training examples of their solution contain images of bolts with multiple types of heads. However, I claim that because the DL architecture of this solution and the type of data it will classify (images of bolts with distinct heads), this solution should use this type of images in its training examples.

[39] By contrast, the second DL modelling methodology exposed by Marino et. al. uses the first validation method. In fact, the model does not need test data for assuring before the testing procedure that it will perform well the generalisation. As the footnote 35 mentions, the architecture of the DL model (a MI architecture) is sufficient for assuring this correct generalisation. This means that the rules the architecture uses for performing the classification decide which parameters and dimensions, independently of the type of bolts inspected, are needed for classifying rails between those that miss bolts and those that do not; the architecture fix these dimensions and parameters; they cannot be changed in distinct classification contexts (hexagonal headed bolts detection, hook headed bolts detection, among others).

solution does not rely on an architecture that relies on fixed parameters for doing the classification (the MI architecture)[40] but on one that requires the use of examples in its training stage and that these examples are composed of data that will be used at the testing stage (images of bolts with multiple types of heads); this architecture is the MFNN. As the second validation method requires that designers construct a hybrid dataset composed of training and testing data, this requirement coincides with the requirement of the MFNN architecture. As mentioned in the previous section, the second validation method is used in inductive-risk-imbalanced contexts, since the errors of the testing procedure are used for defining the possible classification functions of the training and, consequently, the probabilities of each type of errors these functions will have. Thus, using this architecture and, hence, this validation method, shows designers that they are at an inductive-risk-imbalanced validation context[41].

It is fundamental to mention that the MFNN architecture used in the solution proposed by Marino et. al. is an integrated architecture. In fact, this architecture uses two types of data for finding the classification pattern, the training examples dataset composed of testing data (images of bolts with multiple types of heads) and the dataset that will be computed by the model for finding the classification pattern (images of hexagonal-headed bolts). In chapter 3 I have stated that integrated architectures are used in railway systems PdM practices performed at inductive-risk-imbalanced contexts. Therefore, since the possibility of using this architecture indicates designers of DL model for the PdM of railway systems that they are at an inductive-risk-imbalanced validation context, this use of these two types of data is the fundamental feature through which these designers can know that they are at an inductive-risk-balanced validation context.

---

[40] See the previous note.

[41] By contrast, at inductive-risk-balanced validation contexts, designers do not have this possibility of using the architecture of the DL for validating the model. As they use a MI architecture, the training of the model does not depend on the mixing of distinct types of data, but on the selection of a group of features that must be shared by all the possible candidate classification functions. Therefore, the validation cannot be based on constructing hybrid datasets. In the previous section, I have claimed that the first method of validation is the one that fixes the number of dimensions in order to avoid the underfitting and overfitting of the model. Also, I have claimed that this method is used at inductive-risk-balanced contexts, as this fix of the number of dimensions do not alter the probabilities of the errors that might be produced by the classification function. For this reason, as the MI architecture does not allow designers to use the second method, they must use this first validation method. As this method is the one that is used at inductive-risk-balanced contexts, designers can know that, since they are using the MI architecture and, consequently, the first validation method, they are at an inductive-risk-balanced context.

Case 2 - Laser Camera Inspection

Santur, Kakaröse, and Akin (2017) show that contact free inspection systems as the one they propose, in the sense that they use CVSs, can be divided into two groups[42]. On the one hand, there are methods that use normal cameras, i.e., cameras that solely capture colours of the rgb colour model. On the other hand, there are methods that use laser cameras. Lasers are used for capturing the three-dimensional profiles of rails by triangulating the reflected light and, thereby, measuring the height of each of the points that compose the measured rail surface. Santur, Kakaröse, and Akin claim that laser cameras are preferable because detecting faulty rails based on the analysis of rgb images can produce a high rate of false positives. The reason for this is that rail lines are surrounded by residues such as oil and dust. In the second chapter, I have mentioned that the railway visual inspection system proposed by Santur, Kakaröse, and Akin consists in detecting physical alterations in the surfaces of rails, such as scratches. As marks of dust and oil can be registered by the systems as if they were alterations or faults, the system might classify a normal or healthy rail as a faulty one. Since laser cameras calculate the height of each of the points of the rail surface, they ignore changes in the visual characteristics of rails that do not produce changes in these heights, such as the presence of a dust mark. In this way, lasers reduce the number of false positives.

The fact that laser cameras are preferred by maintainers because they reduce the number of false positives shows that, in this second case, maintainers use the first type of validation procedure, i.e., they fix the dimensions that the DL model should employ for generalising its classification procedure. Santur, Kakaröse, and Akin (2017) show that the generalisation procedure done by the fault inspection model must be the same all the time. Conditions such as the dust over the rail track should not be considered as dimensions or parameters for performing the classification procedure. As far as the model can measure the same characteristics of the rail, i.e., the three-dimensional embeddedness of the surfaces of this rail, it will generalise well its classification procedure to

[42] Santur, Kakaröse, and Akin (2017) indicate that railway inspection methods can be divided into manual and automatic. Also, they distinguish in the category of automatic methods two subcategories, contact-based methods, and contact free methods. In contact-based methods, maintainers make a measurement instrument touch the surfaces of the rail and measure a physical quantity, such as friction, or the way ultrasound waves propagate through this rail. Contact free methods are methods based in computer vision systems (CVSs). Consequently, they are cameras, laser or digital, that film rail surfaces in the search for patterns that could represent anomalies in these rails. Santur, Kakaröse, and Akin consider that contact free methods should be preferred, because contact-based methods could damage the rail. The reason for this is that the continuous contact of the measurement instrument with the rail can degrade this second object. Thus, the contact methods can produce false negatives, i.e., classify faulty rails as healthy.

multiple rails. For this reason, classifying a healthy rail as a faulty one (false positive) and classifying a faulty rail as a healthy one (false negative) have the same weight. Both types of errors and the sources that cause them, such as environmental conditions, impede the success of the generalisation. For this reason, in this second case the first type of validation is used. Consequently, as this first type is used at inductive-risk-balanced contexts, the context exposed by Santur, Kakaröse, and Akin is an inductive-risk-balanced validation context.

The criterion for identifying whether the validation of the laser camera inspection system is an inductive-risk-balanced or an inductive-risk-imbalanced context is defined by the type of interaction the inspection system has with the object inspected. On the one hand, cameras that capture just rgb colours do not alter the inspected object but can be confounded by this object. In fact, when the rail has residues, the camera detects faults that it does not have. On the other hand, mechanical inspection systems cannot be confounded but they alter the object (Santur, Kakaröse, and Akin, 2017). Therefore, the systems that guarantees that the model will not produce false positives or false negatives are systems that are able to measure tactile properties, such as the three-dimensional embeddedness of rails, but that, at the same time, do not need tactile interaction. Therefore, the contexts of PdM of railway systems where these types of inspection systems are needed are inductive-risk-balanced contexts.

Santur, Kakaröse, and Akin (2017) indicate that the pooling layer of the CNN used by the DL model they propose is used by the laser inspection system for selecting the features in the image captured. This layer performs a non-linear down-sampling that balances the costs of the datasets produced by the system during the inspection. As mentioned before, the convolution layer fixes the characteristics that the classification model must use for finding the classification pattern. Therefore, these two layers assure that all the features of the images captured have the same weight in the modelling of the classification function and select the features that must be used for finding this function. As the validation method used in inductive-risk-balanced context consists in fixing the parameters that must be used by all the possible classification solutions without considering the weight of these parameters, these two layers perform the validation procedure.

As the inspection system captures an image and isolates the tree-dimensional characteristics of this image for finding the classification pattern, the convolution layer extracts these characteristics, and the pooling layer fixes them in the possible classification solutions. This shows that this joint operation of these two layers is the feature that reveals that this system uses a

separated DL architecture in the validation stage of its DL model. In fact, these two layers isolate the single measurement that the model uses for finding its classification pattern: the three-dimensional properties of the surfaces of the rail tracks.

3. Comparison and contrast

Both cases show that inductive-risk-balanced validation contexts in the use of DL models for the PdM of railway systems are contexts where the inspected characteristics do not change. In the first case, the inspected characteristic is whether the bolt is correctly positioned in the rail track; this characteristic is the same for all types of bolts. In the second case, the inspected characteristic is the three-dimensional embeddedness of the surfaces of the rail. This characteristic is not altered by the environmental conditions. In fact, a laser camera inspection method can detect changes in this embeddedness even if the rail is polluted with dust. For this reason, both cases show that maintainers should evaluate whether the characteristics inspected by the inspection systems are unchangeable characteristics for knowing whether the context where these systems are being validated are inductive-risk-balanced or inductive-risk-imbalanced.

The difference in these cases is that, while in the first one the tactile interaction needed in the inspection process is not a factor that decides whether the validation context is inductive-risk-balanced or inductive-risk-imbalanced, in the second case this factor defines the type of value context. In fact, the criterion for validating the generalisation success of the digital camera inspection system consists in automating the way the model can shift between multiple geometric configurations (hexagonal, hook, among others). This is the only criterion that is used for validating the inspection system. Instead, in the second case, a successful inspection is an inspection that can measure a tactile quality of the railway without having contact with this rail. Therefore, besides analysing a quality that does not change, the inspection system must analyse a tactile quality. Since these tactile qualities can be altered if the inspection system has physical contact with them, a second criterion for a successful generalisation consists in being able to measure these qualities without having physical contact with them. The point of contrast between these two cases is that, while in the first case the type of properties measured by the system are sufficient for determining the validation method and, hence, the type of cost context, in the second case also how the

inspection system access and measures these properties is a fundamental factor in determining the type of cost context.

Chapter 6 - The internal stage of science, inductive risk, and deep learning models

In this chapter, I analyse the internal or training stage of the use of DL. As the other stages, the two types of cost contexts decide which of the two training methods should be used by users of societal applications of DL. Thus, in this chapter I show which cost context corresponds to which method. Next, I argue how this distinction of cost context can be employed for taking design decisions related to the inductive risk problems that arise in the chosen solutions.

1. Deep Learning models and the Bayes method

When constructing societal DL models, designers associate the application of the Bayes method with the *value-costs*[43] of errors (Karaca, 2021:1214; Elkan, 2001). This means that, while true positives and true negatives have utility 1, false positives and negatives vary in their level of utility

[43] It is fundamental to distinguish this definition of the term cost from the one that I introduced before. The first definition of cost I use refers to the impact that an inductive error, a false positive or a false negative, has on human values. For example, in a scientific organisation that seeks to protect consumers safety, falsely claiming that a drug is non-toxic is a false negative that has a greater cost than the false positive claim that states that the drug is toxic. Instead, the second definition of cost that I use refers to the impact that an inductive error has on a sub-value. For example, suppose that the abovementioned drug is for controlling a pandemic disease. The pharmaceutical scientists can develop and distribute the drug swiftly without doing thorough toxicity tests. This will protect the health of the population since it will prevent many people of getting infected. However, this will threat the safety of this population, as there is higher probability for the drug of being toxic. In contrast, if scientists perform thorough toxicity tests, they will assure that the drug is non-toxic, but they will delay the distribution of the drug. This will impact positively the safety of consumers but will negatively impact their health. Thus, the pursuit of health is inverse to the pursuit of safety and scientists must choose which value to prioritise. Suppose that safety and health are sub-values of the value of successfulness. This means that a drug is successful if it is safe, and it has been produced without delay. However, as in this situation pursuing health threats the pursuit of safety and pursuing safety threats the pursuit of health, in order to pursue successfulness, scientists can opt in their tests for prioritising the pursuit of safety and impacting negatively health or prioritising the pursuit of this latter value and impacting negatively the former. In this case, even though false positives and false negatives has the same cost for the value of successfulness, they have distinct costs in the prioritisation. In fact, making a thorough test and reducing the probability of false negatives impacts positively the value of safety but negatively the value of health. In the same way, making a swift testing and reducing the probability of false positives positively impacts the value of health but negatively the value of safety. Thus, the impacts are balanced and none of these errors impacts on a greater way the value of successfulness. However, if scientists perform the test by giving priority to the protection of health, the cost of false negatives is higher than the cost of false positives. For this reason, the cost of an error is different in relation to a value than in relation to a sub-value. The first definition of cost refers to this first impact. The second refers to this second impact. For distinguishing these two definitions I use the term "cost" for referring to the first definition and the term "value-cost" for referring to the second one.

depending on whether a false positive is more impactful than a false negative or vice versa. Therefore, the utility matrix of a DL classification model is the following[44]:

Hypothesis	The statement is true	The statement is false
Classify the statement as true	1 (no error, zero cost)	False positive, cost-2
Classify the statement as false	False negative, cost-1	1 (no error, zero cost)

[44] The Bayes method is a method proposed by the statistician Thomas Bayes for calculating which action to follow based on the *utility* produced by this action and on the *probability* to be true of the assumptions or hypotheses that guide it (Levi, 1962:52-55). For example, a presidential candidate must decide whether to make a huge or just a moderate campaign in a region of the country where elections will be held. She also assumes that, if she makes a moderate campaign but more than 60% of the voters of this region support her, she will win the elections. Additionally, she assumes that, if she makes a huge campaign even if 60% or less of the voters support her, she will win the election. As this candidate must campaign also in other regions and have limited resources, she must decide what to do regarding how to campaign in this first region. Thus, she must analyse which action raises higher levels of utility, or is more useful, given the probability to be true of the hypotheses that guide this action. For doing this, the Bayes method proposes constructing the following matrix:

Hypothesis	More than 60% of the voters prefer her	60% or less of the voters prefer her
Action A: Make a huge campaign	8	9
Action B: Make a moderate campaign	10	7

The highest level of utility is making a moderate campaign given than 60% of the voters prefer her, as she will save the funds of making a huge campaign. For this reason, she assigns a '10' to this combination. The next choice, '9', is making a huge campaign with 60% or less support, since she will guarantee her victory, but she will have to spend the funds of a huge campaign. After this choice, the next one is '8', doing a huge campaign with more than 60% of support, as she will spend unnecessary resources, but she will secure her victory. Finally, '7' is the less useful since her victory is not guaranteed (Levi, 1962:52-55).

If the probability of the first hypothesis (more than 60% of the voters prefer her) is higher than the second (60% or less of the voters prefer her), the candidate must choose action B. Correspondingly, if the second hypothesis has higher probability, the candidate must choose action A. The reasoning process that the candidate applies for calculating the utility of the action is the following: Utility of action A = 8(p) +9(1-p), where p is the probability of the first hypothesis and 1-p is the probability of the second. Thus, supposing that the first hypothesis has 0.5 and the second has 0.25, the formula will be: 8(0.5) +9(1-0.25) (Levi, 1962:52-55).

Isaac Levi (1962:52-55) observes that the Bayes method could be applied to scientific processes. However, he observes that, at this context, it is not possible to assign different utility values to the different combinations, since in science all errors have equal weights. This means that rejecting a true scientific statement has the same weight as accepting a false one. For this reason, the only two values that can be adopted by each of the combinations are 1 or 0, i.e., the scientist makes a correct action (1), or the scientist make an error (0). The matrix will be the next one:

Hypothesis	The statement is true	The statement is false
Decision A - Accept the statement	1	0
Decision B - Reject the statement	0	1

This fact shows that there are two methods for determining the value-cost of an error (Karaca, 2021:12-14). The first method (the *value-cost-sensitive* method) compares the impacts of each error. For doing this comparison, designers must determine which value is more desirable or preferred for the users of the model[45].

In the cases where errors cannot have different value-costs (see note 45), designers must use the second method for assigning costs (the *value-cost-insensitive* method) (Karaca, 2021:1214; Elkan, 2001). This method applies the Bayes method. Hence, the cost[46] of an error is dependent on the probability that this error occurs. For this reason, the probability of the error should be balanced with the cost of the error. Errors that have high probability should have low costs and errors that have low probability should have high costs. In this way, neither of these errors is prioritised.

The cost-sensitive method is used at inductive-risk-imbalanced contexts and the cost-insensitive method is used at inductive-risk-balanced contexts. In the third chapter, I have shown that the contexts where one type of false result has greater impact on human values than the other are inductive-risk-imbalanced contexts. Also, the contexts where errors have balanced impacts or costs are inductive-risk-balanced contexts. Therefore, when a value must be prioritised and designers must use the value-cost-sensitive method for prioritising the false result that has a lower impact on this value, they are at a training context where inductive errors are imbalanced, i.e., an inductive-risk-imbalanced training context. Correspondingly, when values do not need to be prioritised and designers can use the probability of the outcomes for estimating the costs of these

---

[45] For example, in a credit scoring system, if financial security and having high amounts of liquid assets are values prioritised by the bank using the system, the error of giving a loan to a person who has a risky profile has a higher cost than the error of not giving a loan to a person who has a conservative profile.

Nevertheless, there are cases where it is not possible to give different costs to different errors. For instance, in a railway maintenance system, maintainers must take the decision whether to maintain a component or not. This decision is based on whether the component will fail in the next month. Thus, the possible errors are (1) maintaining a component even if it does not need maintenance and (2) omitting the maintenance of a component even if it needs maintenance. Also, the classification hypotheses that are tested by the model are "the component needs maintenance" and "the component does not need maintenance". Moreover, the rule of acceptance is the prognosis that the component will fail in the next month. Therefore, the first error that rejects the first hypothesis without applying the acceptance rule is a false negative. Correspondingly, the second error that rejects the second hypothesis without applying the acceptance rule is a false positive.

Designers could apply the following reasoning and give more weight to the false negative: if the component fails because it is not maintained, the life and safety of the users will be in danger. Also, a railway system where these incidents happen will be closed. However, even if maintaining a component that does not need maintenance in the short term is not problematic, in the long term this unnecessary maintenance might lead to costs overruns and the closure of the system. A system whose maintenance costs surpasses the profits made by this system is unsustainable. Hence, applying unnecessary maintenance might have the same consequence as not maintaining a component that will fail in the following month. This consequence is the closure of the system.

[46] The term of cost used in this sentence refers to the first definition of "cost" introduced previously.

outcomes, i.e., the value-cost-insensitive method, they are at an inductive-risk-balanced training context[47].

In the following section I will analyse how are the inductive-risk-balanced and inductive-risk-imbalanced contexts distinguished at the internal or training stage of the use of DL learning models in the PdM of railway systems. For doing this, I analyse, compare, and contrast the cases of digital camera inspection and of laser camera inspection of railways. The engineers that propose these two models do not specify which *value-cost assigning* method they use for training the model, whether the value-cost-insensitive or the value-cost-sensitive. However, they discuss the distinct types of errors that can be made by the inspection system and how they design the system to decide the value-costs of these errors. In this way, I can establish how they distinguish inductive-risk-balanced and inductive-risk-imbalanced contexts at the training stage of the use of these DL models.

2. Training Deep Learning models, inductive risk, and cost contexts

Case 1 - Digital camera inspection

Marino et. al. (2007:419) claim that their solution alternates between an exhaustive and a jump search for saving computational time and making the inspection process more efficient[48].

[47] Ohnesorge (2020) claims that the value free ideal (VFI) is a philosophical perspective that claims that the role of scientists at the internal stage of science is just assigning probabilities to the distinct results of the scientific process and not to preserve or pursue human values. It could be argued that, as I state that at inductive-risk-imbalanced contexts, the internal stage of science consists in taking methodological decisions based on probabilities, I am defending the VFI. However, this is not the case, since, at this context, scientists are motivated by the values they pursue in this use of probability for assigning costs to errors. I follow the line of thought exposed by Ohnesorge (2020) and Wilholt (2009) that claims that assigning probabilities to scientific outcomes is motivated by values.

[48] According to Marino et. al. (2007:418-419), the distance at which fastening bolts are positioned in a rail is constant for all the bolts. As the bolts fasten the rail to the sleeper, this constant distance is the distance between sleepers. For this reason, the digital camera inspection system consists in inspecting this vertical axis and assure that the bolt is correctly positioned. Programmers establish this distance as a constant parameter. Therefore, the camera always inspects the same parts of the rail in search for bolts. As the camera is mounted in an inspection train, while the train moves, it films the rail at the same vertical axis. Additionally, the camera performs at the beginning of the inspection process an exhaustive search. This means that it films constantly the vertical axis in search for bolts. Nevertheless, as soon as it measures the distance between the first and the second bolt in the same axis, it starts a jump search where it inspects just the segments of the axis that are located at the same distance as that measured between the first and the second bolt. If the system does not find a bolt in its place, it raises an alarm and goes back to the exhaustive search. For this reason, as the system, during the jump search, does not inspect all the rail but the segments it estimates, based on the measured distance between the first and the second bolt, which are the areas where the bolts are located, it is

Consequently, one of the values pursued by this inspection system is *efficiency*. In this system, the false negatives are the outcomes where the system claims that a bolt is in its correct position when it is not. Correspondingly, the false positives are the outcomes where the system claims that the bolt is not in its position when it is. The reason for this is that the hypotheses that the inspection system aims to accept are "the bolt is not correctly positioned (and the rail needs maintenance)" and "the bolt is correctly positioned (and the rail does not need maintenance)". Also, the acceptance rule is the shape detected in the image captured. Therefore, if the system accepts the first hypothesis even if the camera registers a bolt, it produces a false negative. Also, if the system accepts the second hypothesis even if the camera does not register a bolt, it produces a false positive. As the system changes from the jump search to the exhaustive one when it produces a positive result, the probability that the system produces a false positive is higher than the probability that the system produces a false negative. Namely, as the exhaustive search inspects the entire rail, it has lower probability of not registering a missing bolt than the one had by the jump search. Therefore, false negatives are less probable in the exhaustive search than in the jump search. Additionally, as the jump search is applied as far as the system does not produce positive results, the probability of the false negatives of the jump search is conditioned by the probability of the exhaustive search of producing true results. Therefore, the times the system performs a jump search will always be fewer than the times the system performs an exhaustive search; performing a jump search happens just after the performance of an exhaustive search; the opposite will not happen.

As the shift between the exhaustive and the jump search is a *metrological*[49] feature introduced with the aim of achieving higher efficiency, the decrease of the probability of false negatives achieved by this feature is a consequence of this pursuit of efficiency. Marino et. al. (2007:418) claim that missing bolts can reduce safety. Consequently, as false negatives do not register missing bolts, if less false negatives are produced, the maintainers will apply less maintenance actions. Therefore, as the design of the system reduces the probability of false negatives, it reduces the number of maintenance actions. As the system also pursues safety, reducing the probability of false negatives and, consequently, of maintenance actions, is a method

based on a probabilistic or predictive procedure; the system predicts the distance between bolts and inspects these areas.

[49] "Metrological" means that this feature is related to the method through which the system detects or measures a physical feature. In this case, the feature is whether a bolt is correctly positioned in the sleeper. For more information about the relationship between modelling and metrology see Tal (2020), and in particular about the relationship between modelling, metrology and inductive risk see Winsberg (2012).

for optimally pursuing efficiency and safety, i.e., for pursuing safety without negatively impacting the pursuit of efficiency and pursuing efficiency without negatively impacting the pursuit of safety. This shows that this training context is inductive-risk-balanced, as designers, following the value-cost-insensitive method, decrease the probability of one type of error for preserving both safety and efficiency and not for prioritising the pursuit of one of these values.

Inductive-risk-balanced contexts can be distinguished from inductive-risk-imbalanced ones at the training stage by assessing whether a metrological feature of the model can increase or decrease the probability of one of the results. Inductive-risk-balanced contexts are contexts where this artificial increase or decrease, i.e., this change in the probability of an outcome using a specific metrological method or feature as the one described above, is possible. By contrast, inductive-risk-imbalanced contexts are contexts where it is not possible. When models cannot change this probability artificially, the designers of these DL models must assign error costs based on the value they pursue. However, when these models can change this probability artificially, the designers increase the probability of the outcome that favours the value they pursue. Nonetheless, they cannot increase indefinitely this probability, since each of the outcomes has already a preestablished probability defined by the characteristics of the system that are not controlled by the designers. For instance, the probability that the camera malfunctions and increases the probability of false positives is a factor that cannot be controlled by the designers. Therefore, increasing or decreasing probability artificially consists in summing these two types of probability, the artificial and the preestablished. Since this preestablished probability also defines the value that can be obtained by the system, this sum of probabilities is an equilibrium between the values they are related to. In other words, this sum is the optimal pursuit of these values. Consequently, maintainers must assess whether they can change the probability of one of the outcomes of the DL modelling process for knowing whether the training context is an inductive-risk-balanced or an inductive-risk-imbalanced context[50].

[50] It could be objected that designers could also alter the probability of an outcome at an inductive-risk-imbalanced training context. For instance, in a credit scoring system of a risky bank, the users of the system can construct the system in a way that this model increases the probability of false positives, i.e., of customers who are given a credit by the bank even if they do not fulfil the requirements for receiving a credit. This could increase the probability that the bank preserves its value of being financially risk-taking. However, if they do this, they will not need to assign a lower error cost to false positives. In this way, assigning costs based on the values prioritised by the model will not be necessary, as the designers before training already know the classification solution they need for prioritising the pursuit of one of the values. This cannot happen in inductive-risk-imbalanced contexts, since designers cannot increase or decrease the probability of an outcome without knowing first how this artificial change in the probability will affect the prioritisation of one value over another.

This possibility of artificially changing the probability of one of the outcomes of the DL modelling process is revealed by the architecture of this model. Marino et. al. (2007:421-422) indicate that the MLPNC has three layers: input, output, and hidden. Also, they indicate that the same activation function regulates the propagation of the data from the first layer to the second and from the second to the third. This activation function is a sigmoid activation function ($f(x)=1/1+e^{-x}$). For this reason, when finding the classification pattern, the model does not construct the function by combining the error rates linearly but by establishing a threshold that determines whether a certain error costs combination might be used for modelling the pattern or not (see figure 5). As these error rates combinations are created by the alternation between the exhaustive and jump search modes that determine how many false negatives can be produced for each false positive, these modes together with this architecture decide the probability of errors that will be used by the model for finding the classification pattern. Also, since the discovery of this pattern is based on this combination, the measurement used by the DL model for finding this pattern is solely this combination. Therefore, this system uses a separated architecture, and this type of architecture is the characteristic of the model that enables this system to change artificially the probability of one of the outcomes of the DL model and, consequently, to indicate designers that they are at an inductive-risk-balanced training context.

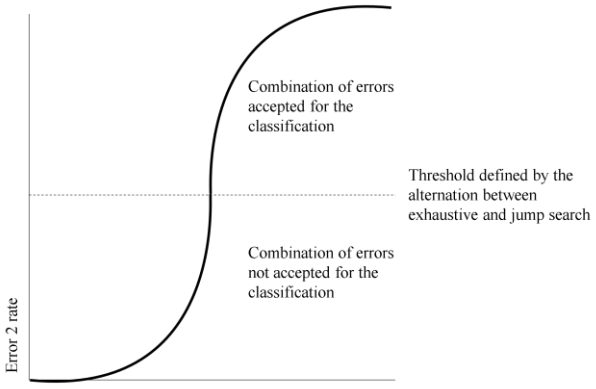

Figure 5: The alternation between exhaustive and jump search modes defines the possible error combinations that the system can produce. Thus, in order to use these combinations to find the classification function that optimally pursues safety and efficiency, designers must define a threshold of combinations that do not prioritise safety or efficiency. This is done through a sigmoid function as the represented by the image. The combinations that are above the threshold can be used for finding the function and, thus, active the MLPNC. In this way, as this architecture is activated by these combinations, it uses them for discovering the classification function based on the data produced by the inspection system.

Case 2 - Laser camera inspection

As in the first case, the false positives are the cases when the system classifies the rail as faulty when it is not. Also, the false negatives are the cases when the system classifies the rail as healthy or normal even if it is faulty. Furthermore, since the normal camera is introduced in the laser camera inspection system with the aim of avoiding false positives, this design choice has the goal of reducing the probability of this type of result. The normal camera tells the system that structural changes of the rail should not be classified as faults. Moreover, the designers of this system show that the laser camera inspection system is not affected by environmental conditions such as dust in the rails and, hence, that, with this type of measurement, the system will not classify healthy rails that have residues of dust that appear to the system as if they were faults as faulty (Santur, Kakaröse and Akin, 2017). Therefore, with this metrological choice, the designers are reducing the probability of false negatives[51]. Since both metrological choices are reducing the probabilities of both types of fault results, these choices are not introduced by designers with the aim of prioritising one result over the other. Therefore, they apply a cost-insensitive value-cost assigning method. In the conceptualisation of inductive-risk-balanced contexts done at the third chapter, I have shown that the contexts where neither false positives nor false negatives are prioritised and, consequently, designers follow a cost-insensitive value-cost assigning method are inductive-risk-balanced contexts. Thus, the training context of this case is inductive-risk-balanced.

In this case designers can increase or decrease the probability of both false positives and false negatives by changing the metrological features of the DL model. Introducing a laser inspection system reduces the probability of false positives and introducing a rgb camera inspection system reduces the probability of false negatives. Santur, Kakaröse, and Akin (2017) claim that the financial cost of laser camera inspection systems is higher than rgb camera inspection. However, they also claim that reducing both false positives and false negatives makes the system more

[51] Santur, Kakaröse, and Akin (2017) claim that the laser camera inspection system consists of a laser camera and a normal rgb camera. The laser, by means of triangulation, can measure the changes in the deepness of the surfaces of the rails and make a profile of how these surfaces are embedded in the three-dimensional space. The rgb camera makes a structural profile of the rail. In this way, maintainers can know whether the changes of the surfaces of these rails are changes caused by the change in the structural profile or are faults such as a crack. For instance, the laser can point to a segment of the rail where a bolt is located. This will cause a change in the three-dimensional profile of the rail. However, this change is not due to a fault, but to the fact that in that segment of the rail there is a bolt. The normal camera aids the system in distinguishing this type of changes from changes caused by faults and, hence, in registering with greater accuracy the changes that occur due to faults.

accurate and, thus, safer. For this reason, I can interpret that their decision for decreasing false positives and false negatives by increasing the cost of the system is a way of prioritising safety over economic efficiency.

Nevertheless, as they do not prioritise false positives over false negatives or false negatives over false positives, they are not prioritising the value of safety. The reason for this is that the occurrence of false negatives is more dangerous than the occurrence of false positives. This fact might present the designers' decisions and goals as contradictory: they want to prioritise safety by reducing the probability of false results, but they are not giving more weight or a higher cost to false negatives, an action that would prioritise safety. They are not contradictory, though. The reason for this is that I can interpret that they do not want to prioritise safety over economic efficiency. Their goal is to find an optimal point where the pursuit of economic efficiency does not negatively impact the pursuit of safety. Consequently, I can state that this case shows that inductive-risk-balanced training contexts are contexts where designers can prioritise one value by increasing or decreasing simultaneously the probability of both false results with the aim not of prioritising the pursuit of this single value but of balancing the pursuit of multiple values.

As in the other case, in this laser camera inspection system the possibility of decreasing the probability of false results by means of a metrological feature is revealed by the architecture of the DL model. As shown in figure 4, this architecture is a CNN composed of a convolutional and a pooling layer, and a set of fully connected layers. I have indicated in the analysis of the validation stage of this case that the first two layers filter the characteristics necessary for the classification and introduce these characteristics into the model. With this process, the model can use just one measurement for discovering the classification pattern. However, it still cannot discover the classification based on the values the system aims to protect. For doing this, it requires an architecture that enables it to assign costs to the possible outcomes of the training process in a way where none of these costs prioritises the pursuit of one of the values over the pursuit of another. As the third layer is a fully connected layer where all the nodes are uniformly connected, this layer does this value-cost-insensitive cost assigning. If this third layer were not fully connected, the model will prioritise those results produced by the nodes that are connected. As the data that enter all the nodes of the layer is produced by the joint work of the rgb camera and laser camera, a layer that is not fully connected might not combine these data in a balanced way and, consequently, will prioritise the pursuit of one value over the pursuit of the others. This shows that just an architecture

that combines in a balanced or uniform way all the data produced by this metrological feature can effectively accomplish the purpose of the introduction of these features: simultaneously regulate the probability of false negatives and false positives in a way where no value is prioritised.

This CNN belongs to the separated architecture category. The reason for this is that the data produced by the metrological feature and processed by the convolution and pooling layers is solely the measurement of the three-dimensional configuration of the surface of the rail track. As the section of this architecture composed with the fully interconnected layers propagates solely the data processed in the pooling layer to the output layer, it also uses this single measurement. Therefore, this propagation is the feature that reveals that the architecture that must be used by the DL model at this training context is from the separated category and, consequently, that the possibility of regulating simultaneously the probabilities of false positives and false negatives with the aim of avoiding prioritising values is created by this architecture.

3. Comparison and contrast

Both cases show that inductive-risk-balanced training contexts can be distinguished from inductive-risk-imbalanced ones because, in the former contexts, the trainers of the DL model and the designers of the inspection system can increase or decrease artificially the probability of false negatives, false positives, or both by changing the metrological features of the system. In the first case, they reduce the probability of false positives by designing a system that alternates between an exhaustive and a jump search. In the second case, they reduce the probability of both false results by combining a laser inspection system with a rgb camera one. Therefore, the criterion maintainers should use for categorising a training context as either an inductive-risk-balanced or as an inductive-risk-imbalanced one consists in determining whether they can increase or decrease the probability of false results by means of changing the metrological features of the system.

Both cases differ in the fact that, while, in the first one, maintainers can just decrease the probability of false negatives, in the second one, maintainers can decrease the probability of both false results. In the first case, designing a system that alternates between an exhaustive and a jump search reduces the probability of false negatives but does not reduce the probability of false positives. By contrast, in the second case, combining a laser system with a rgb camera reduces the

probability of both false positives and false negatives. Thus, whereas, in the first case, having the possibility of increasing or decreasing artificially the probability of one result is sufficient for labelling this training context as an inductive-risk-balanced context, in the second case, there are two criteria: 1) having the possibility of artificially increasing or decreasing the probability of a false result; 2) having the possibility of artificially increasing or decreasing the probability of both false results.

Until this point of the thesis I have responded the second and third sub-questions. The section 3.5-A have shown that managing inductive risk in the societal applications of DL models is done by examining how the errors of the inductive processes performed in these societal applications impact human values and regulating the probabilities of these errors. Also, it shows that DL modelling is a three-stage process (training, validating, and testing) and that each of these stages can be identified with one of the stages of scientific decision-making processes. Moreover, it reveals that identifying the impacts of errors on human values and regulating their probabilities is a process that consists in choosing between two training, two validating, and two testing methods at each of the stages of this modelling process by identifying whether the errors that might be produced at the stage impact human values at the same degree or not. In other words, this process consists in choosing between these two methods at each stage by identifying the value context of these stages. The sections 4.1, 5.1, and 6.1 have shown which are the training, validating, and testing methods that correspond to each one of the multiple training, validating, and testing value contexts:

Stage	Inductive-risk-balanced contexts	Inductive-risk-imbalanced contexts
Training (internal)	Value-cost-insensitive method	Value-cost-sensitive method
Validating (planning)	Fixing dimensions method	Validation dataset method
Testing (external)	ROC	F_1 score

Therefore, since the management of inductive risk consists in distinguishing between rich-data and poor-data contexts at each of the stages of the scientific process where this risk arises, the inductive risk management practices of the societal applications of DL modelling follows a design approach, i.e., relates costs contexts with data contexts at the three stages of this scientific process. Consequently, these practices make this distinction of data-contexts by deciding between these two

types of methods at each of the stages of the scientific process where this risk is to be managed and, consequently, by identifying the value context at each of these stages. In this way, these sections have defined the common framework for assessing the inductive risk that arises in all societal applications of DL (Big Data Analytics) and, thus, answered sub-question 2.

The section 3.5-B has shown that the use of DL modelling in railway systems PdM practices manages inductive risk by monitoring in real-time the impacts of maintenance decisions in human values. Also, it has shown that this monitoring is done by performing this modelling process in the inspection activities of these practices. Furthermore, it has revealed that this performance consists in selecting multiple DL architectures at each of the stages of this process by identifying the type of value context of each of these stages. As this identification depends on assessing how these inspection activities might determine the costs of the inductive errors of these PdM practices, this selection is decided by the study of the relationships between these activities and these costs. Therefore, this section reveals that, in order to explain how the management of inductive risk occurs in the use of DL modelling in the PdM of railway systems, it is necessary to study these relationships. The sections 4.2, 4.3, 5.2, 5.3, 6.2, and 6.3 (case study sections) study these relationships in two cases where DL modelling is used in the inspection activities of railway systems PdM practices. This study confirms that in these activities engineers and maintainers choose the DL architectures to be used at each of the stages of the DL modelling process based on their assessment of the relationships between error costs and the specific characteristics of these inspection activities. Thus, both section 3.5-B and the case study sections reveal that the analysis of the management of inductive risk in railway PdM practices that use DL modelling consists exclusively in assessing the relationships between inspection activities and the costs of errors at each of the stages of this modelling process.

Chapter 7: Conclusion

I have shown that the inductive risk of the predictive maintenance (PdM) of railway systems must be analysed following the same framework and method for analysing the inductive risk of DL and Big Data Analytics. I have shown that the philosophical aspects of the inductive risk problem that pervade the inductive risk of these maintenance practices are the same as those present in the use of this technology and the performance of its modelling process. These are: 1) this inductive risk consists in identifying between four outcomes the pair that negatively impact human values; 2) this inductive risk is global and, consequently, is present at the three stages of the PdM of railway systems and of the use of DL in societal applications; 3) the identification of the risky outcomes depends on identifying whether the context where each of the stages of PdM and of this use of DL is performed is rich-data or poor-data. For this reason, I claim that the philosophical framework for analysing the inductive risk of this use of DL can be used for analysing the inductive risk of the PdM of railway systems.

Moreover, I have established that maintainers and engineers can identify at each stage of the use of DL modelling for the PdM of railway systems whether this stage is being performed at a rich-data or at a poor-data context by distinguishing inductive-risk-balanced contexts from inductive-risk-imbalanced ones at the inspection activities of this use. Additionally, I have shown that this assessment is done by analysing the relationships between data contexts and value contexts revealed by the methodological features of these activities, such as the measurements they make, the fault detection processes they perform, the way the multiple data gathered at these activities are combined, among others. Therefore, I have demonstrated that, if the philosophical framework for analysing DL inductive risk is used for analysing the inductive risk of the PdM of railways systems, this use transforms the philosophical questions that address how to manage the inductive risk of these maintenance practices into design questions. In other words, I have shown that using this framework for analysing the inductive risk that arises in these maintenance practices consists in assessing how the modelling methodologies of the use of DL in this PdM reveal the relations between cost contexts and data contexts at each of the stages of this modelling process.

I think that a fundamental question that arises is how to distinguish between those features of inspection activities that impact the cost context-data context relationships from those that do not have any impact. Eschenbach (2021) claims that the elements of DL societal applications that

impact human values are the elements used for taking high-risk decisions. Additionally, he states that high-risk decisions are those decisions where it is not possible to know prior to taking the decision which human values will be impacted by this decision, i.e., those decisions whose outcomes are *opaque*. As these relationships describe how these activities impact human values, I consider that it is necessary to determine which aspects of inspection are involved in high-risk decisions. For doing this I recommend:

1) *Making a distinction between inspection activities related to the use of a DL model and activities related to the implementation of a DL model.*

Winsberg (2012:127-129) claims that the models used for taking opaque decisions have some elements that can be removed from the model without affecting the effectivity of the model in representing the relationship between the decision and the values it might affect (implementation elements) and other elements that cannot be removed (use elements). Therefore, philosophers that analyse the inductive risk of the PdM of railway systems through the philosophical framework of the inductive risk of DL societal applications should examine which aspects of the inspection activities of these maintenance practices affect this relationship and which aspects do not affect it. Thus, they should remove each aspect and observe whether this representation changes.

2) *Analysing the path dependencies of these inspection activities*

Winsberg (2012:127-129) states that another fundamental feature of the models that represent the relationship between opaque decisions and the impact of these decisions on human values is that which decisions can be taken by modellers for constructing the model depend on other modelling decision taken before. Consequently, philosophers should analyse how the multiple inspection activities are conditionally interconnected or have path dependencies. Thus, they can isolate those activities that do not have any impact on human values and, additionally, that do not determine how other activities that impact human values must be performed.

Word count: 23987

References

Written sources:

Altheide, D., & Schneider, C. (2013). *Qualitative media analysis*. SAGE Publications, Ltd, https://dx.doi.org/10.4135/9781452270043

Alpaydin, E. (2016). Machine Learning: The New AI. The MIT Press.

Bowen, G. A. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, *9*(2), 27-40. https://doi.org/10.3316/QRJ0902027

Chenariyan Nakhaee, M., Hiemstra, D., Stoelinga, M. & van Noort, M. (2019). The Recent Applications of Machine Learning in Rail Track Maintenance: A Survey. In: Collart-Dutilleul, S., Lecomte, T., Romanovsky, A. (eds) Reliability, Safety, and Security of Railway Systems. Modelling, Analysis, Verification, and Certification. RSSRail 2019. *Lecture Notes in Computer Science(), vol 11495*. Springer, Cham. https://doi-org.ezproxy2.utwente.nl/10.1007/978-3-030-18744-6_6

Churchman, C. W. (1948). Statistics, Pragmatics, Induction. *Philosophy of Science*, *15*(3), 249–268. http://www.jstor.org/stable/185088

Costa, J. A. F., & Gonzaga, A. (1996, October). A system for invariant visual pattern recognition using multilayer feedforward neural networks. In *Proc. of Int. Conf. On Signal processing–ICSPAT, Application and Tachnology, Boston MA*

Davari, N., Veloso, B., Costa, G. de A., Pereira, P. M., Ribeiro, R. P., & Gama, J. (2021). A Survey on Data-Driven Predictive Maintenance for the Railway Industry. *Sensors*, *21*(17), 5739. https://doi.org/10.3390/s21175739

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science, 67*(4), 559–579. http://www.jstor.org/stable/188707

Elkan, C.P. (2001). The Foundations of Cost-Sensitive Learning. *IJCAI.*

Erasmus, A., Brunet, T., & Fisher, E. (2021). What is Interpretability?. *Philosophy & technology, 34*(4), 833–862. https://doi.org/10.1007/s13347-020-00435-2

Ghofrani, F., He, Q., Goverde, R.M., & Liu, X. (2018). Recent applications of big data analytics in railway transportation systems: A survey. *Transportation Research Part C-emerging Technologies, 90,* 226-246.

Hossin, M., & Sulaiman, M.N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP), 5*(2), 1-11. https://doi.org/10.5281/zenodo.3557376

Hu, L. & Dai, G. (2022) Estimate remaining useful life for predictive railways maintenance based on LSTM autoencoder. *Neural Comput & Applic*. https://doi-org.ezproxy2.utwente.nl/10.1007/s00521-021-06051-1

Jansen, M. (2021) *Urban AI/ML Models as coupled ethical-epistemic tools of optimization* (Master's thesis). Available from University of Twente Student Thesis (Repository).

Karaca, K. (2021). Values and inductive risk in machine learning modelling: the case of binary classification models. *Euro Jnl Phil Sci 11, 102*. https://doi-org.ezproxy2.utwente.nl/10.1007/s13194-021-00405-1

Levi, I. (1962). On the Seriousness of Mistakes. *Philosophy of Science, 29*(1), 47-65. http://www.jstor.org/stable/185176

Le Nguyen, M.H., Turgis, F., Fayemi, PE. & Bifet, A. (2020). Challenges of Stream Learning for Predictive Maintenance in the Railway Sector. In: Gama, et. al. IoT Streams for Data-Driven

Predictive Maintenance and IoT, Edge, and Mobile for Embedded Machine Learning. ITEM IoT Streams 2020 2020. *Communications in Computer and Information Science, vol 1325*. Springer, Cham. https://doi-org.ezproxy2.utwente.nl/10.1007/978-3-030-66770-2_2

Marino, F., Distante, A., Mazzeo, P. L., & Stella, E., (2007). A Real-Time Visual Inspection System for Railway Maintenance: Automatic Hexagonal-Headed Bolts Detection. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 37*(3), 418-428. doi: 10.1109/TSMCC.2007.893278.

Ohnesorge, M. (2020). The Limits of Conventional Justification: Inductive Risk and Industry Bias Beyond Conventionalism. *Front. Res. Metr. Anal.* 5:599506. doi: 10.3389/frma.2020.599506

Russell, A. & Vinsel, L. (2016). Hail the Maintainers. Edited by Sam Haselby. *Aeon Media Group 2012-2022*. https://aeon.co/essays/innovation-is-overvalued-maintenance-often-matters-more

Santur, Y., Karaköse, M. & Akin, E. (2017). A new rail inspection method based on deep learning using laser cameras. In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 2017, pp. 1-6, doi: 10.1109/IDAP.2017.8090245.

Serradilla, O., Zugasti, E., Rodriguez, J. & Zurutuza, U. (2022) Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects. *Appl Intell 52*, 10934–10964. https://doi-org.ezproxy2.utwente.nl/10.1007/s10489-021-03004-y

Schiele, H., Bos - Nehles, A. C., Stegmaier, P., Delke, V., & Torn, I. A. R. (2021). Interpreting the industry 4.0 future: technology, business, society and people. *Journal of business strategy*. https://doi.org/10.1108/JBS-08-2020-0181

Stegenga, J. (2017). Medical Nihilism. OUP Oxford.

Stella, E., Mazzeo, P.L., Nitti, M., Cicirelli, G., Distante, A., & D'orazio, T. (2002). Visual recognition of missing fastening elements for railroad maintenance. *Proceedings. The IEEE 5th International Conference on Intelligent Transportation Systems*, 94-99.

Society for Technical Communication (2019, February 10). *"Defining Technical Communication".* Society for Technical Communication. https://www.stc.org/about-stc/defining-technical-communication/

Tal, E. (2020). Measurement in Science. In: E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition). Stanford University URL = <https://plato.stanford.edu/archives/fall2020/entries/measurement-science/>.

Weinmann, M. (2013). Visual Features—From Early Concepts to Modern Computer Vision. In: Farinella, G., Battiato, S., Cipolla, R. (eds) Advanced Topics in Computer Vision. Advances in Computer Vision and Pattern Recognition. Springer, London. https://doi-org.ezproxy2.utwente.nl/10.1007/978-1-4471-5520-1_1

Wilholt, T. (2009). Bias and Values in Scientific Research. Studies in *History and Philosophy of Science Part A* 40(1). 92-101 https://doi.org/10.1016/j.shpsa.2008.12.005

Winsberg, E. (2012). Values and uncertainties in the predictions of global climate models. *Kennedy Institute of Ethics journal*, *22*(2), 111–137. https://doi.org/10.1353/ken.2012.0008

von Eschenbach, W.J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philos. Technol. 34*, 1607-1622. https://doi-org.ezproxy2.utwente.nl/10.1007/s13347-021-00477-0

Xie, J., Huang, J., Zeng, C., Jiang, S.-H., & Podlich, N. (2020). Systematic Literature Review on Data-Driven Models for Predictive Maintenance of Railway Track: *Implications in Geotechnical Engineering. Geosciences, 10*(11), 425. https://doi.org/10.3390/geosciences10110425

Xu, T., Tang, T., Wang, H. & Yuan, T. (2013). Risk-Based Predictive Maintenance for Safety-Critical Systems by Using Probabilistic Inference. *Mathematical Problems in Engineering*, vol. 2013, Article ID 947104, 9 pages, https://doi.org/10.1155/2013/947104

Wang, H., Zheng, H. (2013). Model Validation, Machine Learning. In: Dubitzky, W., Wolkenhauer, O., Cho, KH., Yokota, H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY. https://doi-org.ezproxy2.utwente.nl/10.1007/978-1-4419-9863-7_233

Zainal, Z. (2017). Case Study As a Research Method. *Jurnal Kemanusiaan, 5*(1). Retrieved from https://jurnalkemanusiaan.utm.my/index.php/kemanusiaan/article/view/165

Zaharah Allah Bukhsh, Aaqib Saaed, & Irina Stipanovic. (2018, April 17). A machine learning approach for maintenance prediction of railway assets. Transport Research Arena TRA (TRA), Vienna. https://doi.org/10.5281/zenodo.3381949

Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy and Technology*, *34*(2), 265-288. https://doi.org/10.1007/s13347-019-00382-7

Figures:

Figure 1: Fraga-Lamas, P., Fernández-Caramés & Castedo, L. (2017). Railway communication scenarios [Diagram and photography]. In Fraga-Lamas, P., Fernández-Caramés, T. M., & Castedo, L. (2017). Towards the Internet of Smart Trains: A Review on Industrial IoT-Connected Railways. *Sensors (Basel, Switzerland), 17*(6), 1457. https://doi.org/10.3390/s17061457

Figure 2: Huang, M., Liu, Z., & Tao, Y. (2020). Topology diagram of predictive maintenance of mechanical equipment based on IoT [Diagram]. In Huang, M., Liu, Z., & Tao, Y. (2020). Mechanical fault diagnosis and prediction in IoT based on multi-source sensing data fusion. *Simul. Model. Pract. Theory, 102*, 101981.

Figure 3: Santur, Y., Kakaröse, M. & Akin, E. (2017). A) Rail surface; b) Some 3d Rail profiles for learning in dataset; c) Merged 3d rail profile; d) Faulty rail profile [Album of pictures]. In Santur, Y., Karaköse, M. & Akin, E. (2017). A new rail inspection method based on deep learning using laser cameras. In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 2017, pp. 1-6, doi: 10.1109/IDAP.2017.8090245.

Figure 4: Santur, Y., Kakaröse, M. & Akin, E. (2017). CNN model for deep learning based rail inspection. [Illustration] In Santur, Y., Karaköse, M. & Akin, E. (2017). A new rail inspection method based on deep learning using laser cameras. In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 2017, pp. 1-6, doi: 10.1109/IDAP.2017.8090245.