Non-Functional Requirements for Machine Learning: An Exploration of System Scope and Interest

Khan Mohammad Habibullah
Gregory Gay
Jennifer Horkoff
{khan.mohammad.habibullah,jennifer.horkoff}@gu.se
greg@greggay.com
Chalmers | University of Gothenburg
Gothenburg, Sweden

ABSTRACT

Systems that rely on Machine Learning (ML systems) have differing demands on system quality compared to traditional systems. Such quality demands, known as non-functional requirements (NFRs), may differ in their definition, scope, and importance from NFRs for traditional systems. Despite the importance of NFRs for ML systems, our understanding of their definitions and scope—and of the extent of existing research in each NFR—is lacking compared to our understanding in traditional domains.

Building on an investigation into importance and treatment of ML system NFRs in industry, we make three contributions towards narrowing this gap: (1) we present clusters of ML system NFRs based on shared characteristics, (2) we use Scopus search results—as well as inter-coder reliability on a sample of NFRs—to estimate the number of relevant studies on a subset of the NFRs, and (3), we use our initial reading of titles and abstracts in each sample to define the scope of NFRs over parts of the system (e.g., training data, ML model, or other system elements). These initial findings form the groundwork for future research in this emerging domain.

CCS CONCEPTS

• **Software and its engineering** → **Extra-functional properties**; **Requirements analysis**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Non-Functional Requirements, Machine Learning, Machine Learning Systems, Requirements Engineering

ACM Reference Format:
Khan Mohammad Habibullah, Gregory Gay, and Jennifer Horkoff. 2022. Non-Functional Requirements for Machine Learning: An Exploration of System Scope and Interest. In *CAIN '22: 1st International Conference on AI Engineering, May 21–22, 2022, Pittsburgh, PA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/1122445.1122456

CAIN '22, May 21–22, 2022, Pittsburgh, PA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Machine Learning (ML) is increasingly being used in decision making and prediction applications that influence many aspects of our lives. Complex systems, referred to as ML systems, use ML to deliver critical functionality. Such systems demand high computational capabilities (often based on GPUs), process a large amount of data, and utilize complex non-deterministic algorithms [12]. Therefore, ensuring the quality of such systems is potentially more expensive and effort-intensive than traditional systems.

A thorough requirements engineering (RE) process is necessary to ensure the quality of complex systems. Requirements imposed on system quality are known as non-functional requirements (NFRs) [7], and are expressed over different assets of quality [10]. For a ML system, one might easily imagine constraints imposed over attributes such as fairness [15], transparency [2], privacy [4], security [21], or safety [6].

Although significant research has been devoted to NFRs, significant challenges remain for modern system development [14]. Even for traditional systems, NFRs are difficult and challenging to express explicitly [9], and even more difficult to verify or validate [22]. Such challenges compound for ML systems, and the identification, definition, and measurement of NFRs for ML systems has emerged as a critical problem to solve [11, 12].

Much of our accumulated knowledge concerning NFRs may be no longer relevant when dealing with ML systems, due to their complex and non-deterministic nature. Some NFRs, such as fairness, explainability, and privacy become more important. Others, such as usability or interoperability, may become less important [11, 15, 23]. New NFRs, such as retrainability, emerge. Moreover, the meaning and interpretation of NFRs for ML systems may differ from traditional systems and may not yet be well understood [1]. To date, there has been little research on NFRs for ML systems [12].

Further, "ML" is not one monolithic entity, but can be considered at different levels of granularity within a larger system [24]. When imposing NFRs over an ML system, some NFRs may apply to the algorithm that performs the learning task, while others may apply over the training data used as the basis for such decision making or over the model trained using that data. Still others may apply over the results of applying that model, or over the broader ML system that acts on those results. Therefore, the scope of consideration for NFRs (i.e., the scope of identification, definition, and measurement) for ML systems is a complex and not-yet-solved problem.

A recent interview study, conducted by Habibullah and Horkoff, examined treatment of NFRs for ML-systems in industry, and reported challenges of identifying, defining and measuring NFRs [11]. Addressing these challenges will require (1) **a detailed understanding of the definition and scope** of each NFR in a ML system context, and (2), **an examination of past research** on each of these NFRs as applied to the development of ML systems. These needs are intertwined. To date, there has been no systematic literature reviews or other secondary studies on ML system NFRs. However, performing such a study requires a clear answer to questions of scope to proceed effectively.

In this study, we perform an exploratory study of the treatment of certain NFRs for ML systems in research literature[1]. Our goals in this study are to (1) gain an approximate idea of the extent to which select NFRs have been studied by researchers, and (2), perform an initial clarification of the scope of these NFRs for ML systems.

As a starting point, we have taken into account the NFRs identified as important in the interview study [11]. Using this set of NFRs, we have made the following contributions:

- We divide these NFRs into a set of clusters based on shared attributes of their definitions. This enables better understanding of which NFRs could be considered in conjunctions with each other. For example, researchers could develop studies focused on particular clusters, and practitioners may consider defining system quality over related NFRs.
- We identify an upper limit on the number of relevant publications that exist for each NFR in the Scopus literature database. Our initial estimation shows that some NFRs, such as security or transparency, have received significant focus. Therefore, we focus on NFRs that have received less attention (e.g., maintainability or testability). To gain a finer estimation for these NFRs, we selected a subset of the NFRs and examine the titles and abstracts of 50-100 publications for each. Based on this sample, we estimate the number of relevant publications on each of the selected NFRs. This estimation can be used by researchers to identify which NFRs have attracted greater interest, and which could benefit from more attention in the future. It also enables scoping of further secondary studies.
- Based on inspection of the titles and abstracts of these samples, we perform an exploratory scoping of the selected NFRs in terms of which elements of the system they can be defined over (e.g., training data, ML algorithm, or ML model). This scoping brings further clarity to the specific definitions and treatment of these NFRs, which can benefit future research and practice on each.

Our study, while exploratory in nature, is intended to open new opportunities for future research in NFRs for ML systems. We hope to set the groundwork for future studies by clarifying the scoping and definitions for these NFRs, identifying connections between NFRs, and gaining an approximate idea of past interest in these NFRs. Our results can allow researchers to plan future studies and to identify NFRs that have not received sufficient attention. They also help enable engineers to identify which NFRs to consider in conjunction with others of interest, and to think critically about how NFRs apply to different facets of the system-under-development.

[1]While little work has been conducted on the topic of NFRs for ML systems, there is certainly relevant research on individual quality attributes, such as fairness or security.

2 BACKGROUND AND RELATED WORK

In this section we give an overview of necessary background and relevant related work. We give a brief overview of research on NFRs, RE for ML, then the emerging topic of NFRs for ML systems. We describe how ML is part of a large system and how this may effect NFR definitions for ML systems.

2.1 NFRs for Traditional Systems

NFRs are considered essential for ensuring the quality of software, but there are no agreed guidelines on how and when NFRs should be elicited, defined, documented, and validated [10]. Moreover, there is no consensus in the requirements engineering (RE) community regarding which step of the RE process NFRs should be considered and applied [7]. Significant research has been devoted to NFRs in RE, e.g., Doerr et al. applied a systematic, experience-based, method to elicit, document, and analyze NFRs with the objective of creating a sufficient set of traceable and measurable NFRs [8]. While most work such as this focuses on NFRs for traditional software, we are focused on NFRs specifically for systems that make use of ML to deliver functionality. Although many researchers have studied NFRs for traditional systems, very few studies to date have focused on NFRs for ML systems.

2.2 Requirement Engineering for ML Systems

Although there are approaches on how to use ML to improve RE tasks, there has not been extensive research on RE for ML systems [25]. Engineering of ML systems requires different and novel approaches due to their unpredictable nature and differences in their development process. It is crucial to clearly identify and define these differences, in order to offer tailored practices [13]. For traditional systems, activities related to requirements analysis and specification are often performed in the early phases of development, with requirements used downstream as part of design, implementation and verification. However, the activity flow often differs for ML systems due to their reliance on data and the unpredictability of ML results. Upfront problem definition for ML systems can be difficult, as building a clear definition of the problem often requires iterative exploration of data and processes—more so than in typical systems. As such, RE for ML systems has many unknown and unexplored areas, including an understanding of how NFRs differ for such systems.

2.3 NFRs for ML Systems

Horkoff discussed challenges and research directions for NFRs for ML systems [12]. Some of our knowledge about NFRs for traditional systems may no longer be applicable due to the non-deterministic behavior of ML-enabled systems, as well as due to additional performance demands imposed by the need to process and act on large volumes of data. Furthermore, some NFRs become more important (e.g., explainability), some become less relevant (e.g., modularity), there are differing trade-offs between NFRs (e.g., increased security often causes decreased usefulness), and there is no unified collection and consideration of NFRs for ML-enabled systems. Horkoff

defines research direction for NFRs for ML systems, including exploring and defining NFRs for ML, as well as reinterpreting and redefining NFRs that already exist for traditional systems.

In a recent interview study, Habibullah and Horkoff examined challenges regarding NFRs for ML in industry by identifying examples of the identification and measurement of NFRs and examining the importance that practitioners place on NFRs for ML [11]. In contrast, in our study, we build on these results while also exploring the treatment of such NFRs in existing academic research.

The results of the interview study found that most NFRs as defined for traditional software are still relevant for ML-enabled systems. Some NFRs, such as flexibility, efficiency, usability, portability, reusability, and usability were identified as less important by some interviewees. However, they were still considered important by other interviewees. The NFRs identified in the interview study are listed in Table 1. We have defined each NFR, often in an ML context, based on both our experience and related literature, such as research papers, websites, blogs, and forums.

In addition to identifying relevant NFRs, Habibullah and Horkoff reported several gaps to address in future work, including the identification, definition, and perceptions of NFRs in an ML-context [11]. They stated that there are gaps in the definitions and methods to address the scope of NFR coverage in an ML system. In this work, we build on the results of the interview study by beginning to explore the coverage of NFRs in research literature.

2.4 ML as Part of a Larger Software System

In a ML system, the "ML" is a small part of a larger system [24]. In traditional systems, NFRs can be identified over the whole system or elements of the system. In an ML context, NFRs can also be seen to have varying scope (i.e., can be defined over different parts of the system). However, for ML systems, these elements may differ from traditional systems, and the differing nature of these elements may lead to a differing understanding of relevant NFRs. In our preliminary NFR definitions in Table 1, we have sometimes defined NFRs in ML terms, referencing the ML model or data. However, we see that this is not done consistently, and not all potential elemtents of the system are considered. In an effort to improve how NFRs for ML systems are defined, we explore the idea of NFR "scope" further as part of this study.

3 METHODOLOGY

Though NFRs for traditional software are fairly well-understood, there are still gaps in our foundational knowledge on NFRs for ML systems. We are eager to learn which NFRs for ML systems have been explored by other researchers and which are yet to be investigated heavily. We also want to learn how NFRs for ML are perceived by other researchers so that the definitions and scopes of such NFRs can be refined.

Hence, we have performed an exploratory study aimed at at establishing an initial scoping of the treatment of NFRs for ML and an initial estimation of the level of research that has been conducted on these NFRs. A systematic mapping study is primarily concerned with structuring a research area [17]. As we are performing an initial exploration of the scope of NFRs for ML systems, we have adapted the systematic mapping study concept for our purposes.

3.1 Research Questions

There is lack of knowledge on NFRs for ML-enabled systems, and the perception and current treatment of NFRs have yet to be thoroughly explored. Hence, our goals in this study are to (1) gain an approximate idea of the extent to which select NFRs have been studied by researchers, and (2), perform an initial clarification of the scope of these NFRs for ML systems. We perform this clarification by grouping NFRs into clusters based on shared characteristics, then defining which elements of a ML system (e.g., training data, model, ML algorithm) that each NFR can be established over using an initial sampling of the research literature.

Specifically, we address the following research questions:

- RQ1 Can the ML system NFRs be grouped into a small number of clusters based on shared characteristics?
- RQ2 Which NFRs have received the most—or least—attention in existing research literature?
- RQ3 Over which elements of an ML system can individual NFRs be defined?

We performed the following to answer our research questions:

(1) Based on the definitions in Table 1, as well as our understanding of each NFR from past experience and literature, we grouped the NFRs into a small number of clusters based on their shared characteristics (Sec. 3.2).
(2) We performed the initial stages of a mapping study in order to gain a rough approximation of the how much research exists on each NFR—focusing on those NFRs that have been least investigated or belong to two particular clusters of interest (Sec. 3.3).
    (a) We retrieve the number of publications from Scopus for each NFR based on a search string that we formulated. This gives an upper limit on the number of potentially relevant publications (Section 3.3.1).
    (b) To refine this estimation, we inspected the abstract and title of a sample of the retrieved publications for a subset of the NFRs (Section 3.3.2). We measure intercoder reliability (ICR) on this sample, and use the final agreed-upon set of relevant papers to gain an approximate estimate of the number of relevant papers that exist for this sample (Section 3.3.3).
(3) Based on knowledge gained from reading the titles and abstracts from these samples and past experience, we identify which elements of the system that these NFRs should be defined and measured over (Section 3.4).

3.2 NFR Clustering

In Table 1, we listed the NFRs found to be important in the interview study [11]. For each, we have defined them based both on our past experience and based on their treatment in a small sampling of research papers, websites, blogs, and forums. Based on these definitions, we are interested in grouping these NFRs into a small number of clusters, where each cluster contains NFRs that have similar meaning or purpose. Researchers can use these clusters to identify which NFRs may be related and able to collectively determine the quality of a system. Researchers could also perform secondary studies on particular clusters of NFRs. Developers can

Table 1: Important NFRs for ML systems, identified in [11].

NFRs	Definition
Accuracy	The number of correctly predicted data points out of all the data points.
Adaptability	The ability of a system to work well in different but related contexts.
Bias	A phenomenon that occurs when an algorithm produces results that are systematically prejudiced due to erroneous assumptions in the ML process.
Completeness	An indication of the comprehensiveness of available data, as a proportion of the entire data set, to address specific information requirements.
Complexity	When a system or solution has many components, interrelations or interactions, and is difficult to understand.
Consistency	A series of measurements of the same project carried out by different raters using the same method should produce similar results.
Correctness	The output of the system matches the expectations outlined in the requirements, and the system operates without failure.
Domain Adaptation	The ability of a model trained on a source domain to be used in a different—but related—domain.
Efficiency	The ability to accomplish something with minimal time and effort.
Ethics	Concerned with adding or ensuring moral behaviors.
Explainability	The extent to which the internal mechanics of ML-enabled system can be explained in human terms.
Fairness	The ability of a system to operate in a fair and unbiased manner
Fault Tolerance	The ability of a system to continue operating without interruption when one or more of its components fail.
Flexibility	The ability of a system to react to changing demands or conditions.
Integrity	The ability to ensure that data is real, accurate and safeguarded from unauthorised modification.
Interpretability	The extraction of relevant knowledge from a model concerning relationships either contained in data or learned by the model
Interoperability	The ability for two systems to communicate effectively
Justifiability	The ability to be show the output of an ML-enabled system to be right or reasonable.
Maintainability	The ease with which a system or component can be modified to correct faults, improve performance or other attributes, or adapt to a changed environment
Performance	The ability of a system to perform actions within defined time or throughput bounds.
Portability	The ability to transfer a system or element of a system from one environment to another.
Privacy	An algorithm is private if an observer examining the output is not able to determine whether a specific individual's information was used in the computation.
Reliability	The probability of the software performing without failure for a specific number of uses or amount of time.
Repeatability	The variation in measurements taken by a single instrument or person under the same conditions.
Retrainability	The ability to re-run the process that generated the previously selected model on a new training set of data.
Reproducibility	One can repeatedly run your algorithm on certain datasets and obtain the same (or similar) results.
Reusability	The ability of reusing the whole or the greater part of the system component for similar but different purpose.
Safety	The absence of failures or conditions that render a system dangerous
Scalability	The ability to increase or decrease the capacity of the system in response to changing demands.
Security	Security measures ensure a system's safety against espionage or sabotage.
Testability	The ability of the system to to support testing by offering relevant information or ensuring the visibility of failures.
Transparency	The extent to which a human user can infer why the system made a particular decision or produced a particular externally-visible behaviour.
Traceability	The ability to trace work items across the development lifecycle.
Trust	A trusted system is a system that is relied upon to a specified extent to enforce a specified security, or a security policy.
Usability	How effectively users can learn and use a system.

also use these clusters to identify which NFRs may be relevant to their particular needs or system-under-development.

As a starting point, we have created these clusters primarily through discussion of the NFRs and their definitions. During a series of meetings, we read and interpreted the definitions and debated their meaning. We then discussed and decided which cluster to assign an NFR to. We have placed NFRs in clusters if they are a similar purpose within system development or could be measured in a similar manner.

For example, the explainability of a ML system refers to the extent to which its internal mechanics can be explained in human terms. Transparency refers to the ability of the system to clarify the reasoning for its decisions to a human user. These NFRs differ in their exact meaning and assessment, but are both key elements in ensuring that ML systems operate in a clear and reasonable manner. Therefore, both should reside in the same cluster.

We also created a separate cluster for those NFRs that could not be put into any of the other clusters, as they lacked any shared characteristics with the NFRs in other clusters.

Our goal at this stage is not to create a formal hierarchy, as exists for NFRs for traditional systems [3]. Rather, our interest is in creating a lightweight organizational structure for use in understanding the scoping and definition of NFRs for ML systems. We explain the resulting clusters and their contents in Section 4.1.

3.3 Publication Volume Estimation

In this section, we describe our strategy for estimating the number of research papers for certain NFRs.

3.3.1 Initial Paper Search. We performed a database search—including all publications up to September 2021—in order to identify the research papers that may be relevant for each NFR. We selected Scopus, a meta-database, which includes research papers from peer-reviewed journals and conferences from multiple publishers such as IEEE, ACM, and Elsevier. Scopus is considered one of the most representative and rich in content for Software Engineering research and is used in many secondary studies [16].

We identified relevant search terms and developed search strings for the database search. We first derived the major terms (e.g., machine learning, non-functional requirements). Then, we identified synonyms or alternative spelling for the major terms from related literature, and based on our discussions. We also split major terms into more specific and clear terms. For example, we split the general term "non-functional requirements" into strings based on specific NFRs. Finally, we concatenated these terms to form search strings.

We apply one search string per NFR. The string includes that NFR, as well as terms related to machine learning: **("machine learning" OR "supervised learning" OR "unsupervised learning" OR "reinforcement learning" OR "deep learning")**.

For example, to identify papers on interoperability, we have used the search string: **("machine learning" OR "supervised**

learning" OR "unsupervised learning" OR "reinforcement learning" OR "deep learning") AND ("interoperability"). As a second example, to identify papers related to usability, we have used the string: **("machine learning" OR "supervised learning" OR "unsupervised learning" OR "reinforcement learning" OR "deep learning") AND ("usability")**.

The number of papers found from this step give an upper limit on the number of relevant publications. Not all of these publications are likely to be relevant, as they may not relate to the use of such properties as NFRs for a ML system. For example, several of the results for maintainability described work which used ML to predict maintainability of another system, rather than focusing on maintainability of an ML system. Therefore, in the next step, we used a sample of publications to gain a finer estimation of the number of relevant publications for a subset of the NFRs.

3.3.2 NFR Selection. This upper limit gives some indication of the research interest in each NFR. To gain a clearer estimation of the percentage of those publications that are relevant, we have chosen to focus on a subset of the list of NFRs. Some NFRs, such as performance or security, have already received significant attention from the research community. We would recommend that future secondary studies focus specifically on these topics. We have instead focused on those NFRs that have received less focus from researchers, including those with a lower number of publications as well as those that we identified as being part of two clusters of interest (the "other" cluster and a cluster centered around tailoring a system to different environments).

We created a list of the number of publications found in the search results for each NFR. At first, we sorted the NFRs based on the number of publications, in decreasing order. We then excluded those NFRs that have more than 1,200 search results. For example, we excluded accuracy, as the number of retrieved papers was more than the threshold. Based on this threshold, we excluded 16 NFRs.

We then took into account which cluster we assigned each NFR to. If an NFR has more search results than the threshold but falls into the two clusters that we selected for initial inspection, then we included that NFR for consideration. As a result, we reincorporated usability and flexibility into our estimation, as those NFRs fall into these two clusters even though those have more search results than the threshold. We perform a more detailed analysis on 20 NFRs.

3.3.3 Estimating the Number of Relevant Papers for Selected NFRs. We estimate the number of relevant publications for each selected NFR by inspecting the titles and abstracts of a sample of 50 papers. We read the title and abstract of each publication and use inclusion and exclusion criteria to filter these publications, marking them as relevant or irrelevant. Each author determined the relevancy of each paper independently. We then discussed each disagreement in a meeting, using our criteria, and formed a final list.

Inclusion and Exclusion Criteria: To be **included**, the publication must meet the following criteria:

- The publication must discuss an NFR for from Table 1.
- The publication must focus on the definition, identification, measurement, or challenges of a NFR for a ML system, or for an element of the system (e.g., the model).
- The publication must have been published in a peer-reviewed journal, conference, or workshop.
- The full text of the publication must be accessible and written in English.

Publications that meet the following criteria are **excluded**:

- The publication is focused on topics other than NFRs for ML systems. This includes publications where ML is used to measure, improve, or predict a NFR. For example, the authors used ML to classify requirements into different NFRs [18]. In such a case, the publication is not relevant for examining how such an NFR affects the development of a ML system.
- The publication simply uses the NFR as an evaluation criteria, but does not discuss or describe the use of the NFR during system development. For example, if an author uses completeness as part of their evaluation of the results of a system, but the actual research has no relation to improving the completeness of a ML system, then it is excluded.
- The publication was not written in English, not peer-reviewed, or lacks an available full text.
- Editorials, abstracts, book chapters, workshop summaries, poster sessions, prefaces, article summaries, interviews, news, reviews, comments, news, reviews, tutorials, panels, and discussions are excluded.

Inter-coder Reliability: Following the process of individually reviewing 50 papers for selected NFRs, we calculated our agreement using Fleiss' kappa, a statistical measure for assessing ICR between a fixed number of raters. In some cases where the ICR was low, or where there were significant disagreements, we repeated the sampling process for a second set of 50 papers. In such cases, it was hoped that we could clarify our shared definition and estimation of the scope of the NFR. If the ICR either increased or stayed the same, this served to increase confidence in our understanding.

The final list of relevant papers, after discussion, gives an indication of the number of publications that may be relevant from that initial set retrieved from Scopus. This, in turn, offers an indication of research interest in the NFR.

Estimating the Number of Publications: We counted the number of publications that were deemed relevant from the first—and, in some cases, second—sample for each selected NFR after our discussions. We used these counts along with the total number of papers found by Scopus to estimate the total number of included papers. This estimation is calculated by simply multiplying the total number of publications by the percentage of the sample that was deemed relevant.

For example, we found an upper limit of 851 publications for the transparency NFR. After screening 50 publications, we agreed to include 44 (88%). Extending to the full set of 851 papers, we estimate that 749 publications will actually be relevant. As a second example, we identified 214 publications for traceability. In this case, we sampled 100 publications, and decided that 10 were relevant (10%). Therefore, approximately 21 of the 214 are expected to be relevant to the treatment of the property as a NFR for ML systems.

We repeated this calculation for the rest of the selected NFRs, producing an estimation of the number of relevant papers for each.

This is still a rough approximation of past research interest, but it is sufficient to provide an initial portrait of the field and to refine our own definitions and ideas regarding scope.

3.4 NFR Scope Determination

In order to clearly define or measure the attainment of an NFR, it must be understood exactly how the NFR applies to the system. This determination requires understanding whether a NFR relates to the system as a whole, or perhaps to a lower level of granularity within the system. In the case of a ML system, an NFR may be defined and measured over different aspects of the ML application. For example, an NFR may apply differently when we discuss the training data, the algorithm that uses the training data to build a model, or to the model trained on that data.

Therefore, we have first determined which elements of a ML system are particularly relevant when we discuss the NFRs for such systems. We then used our existing definitions, past experience, and the titles and abstracts of the relevant studies examined in the previous step in order to determine to which of these elements each NFR was applicable. In a series of meetings, we discussed each NFR in relation to these system elements. In each case, we made a determination by coming to an agreement and discussing any cases where we disagreed—generally by identifying an example of how that NFR is applied to that element. For example, repeatability refers to the level of variability in the behavior of the system. Repeatibility is a property of the results—or of the system as a whole—rather then a property of the model, algorithm, or training data. It is the results that vary, not the model itself.

This scoping is intended as a starting point for establishing detailed definitions for each NFR in an ML system context. We present the results of this process in Sec. 4.3.

4 RESULTS

We describe the results of clustering related NFRs in Sec. 4.1, our estimation of research interest in each NFR in Sec. 4.2, and our examination of the elements of a ML system that each NFR can be defined over in Sec. 4.3.

4.1 NFR Clustering Results (RQ1)

We are interested in clustering the list of relevant NFRs from [11], presented in 1, based on which NFRs share a close association. In order to determine the clusters, we followed the process described in Sec. 3.2. We were able to create six different clusters, where each cluster includes the NFRs that share similar properties and purposes. For example, after analyzing their definitions, we found that ethics, bias, and fairness shares similar meanings and serve similar purposes. Therefore, we put these three NFRs into the same cluster. We also created another cluster for those NFRs who meaning and purpose do not match with those in any other clusters. Usability, completeness, maintainability, traceability, testability, and retrainability are included in this cluster because they do not fit into the six identified clusters.

The clusters are presented in Fig. 1. They are as follows:

- Cluster 1 includes NFRs that are related to assessing the functional correctness of ML systems and aspects of correctness. This includes the core correctness, as well as assessment of correctness (e.g., accuracy) and variance (e.g., reliability, consistency).
- Cluster 2 contains NFRs related to understanding the internal decisions or results of applying ML (e.g., transparency, explainability).
- The NFRs related to ethical aspects of ML systems, such as fairness and bias, form cluster 3.
- NFRs related to the performance (e.g., speed) of an ML system are contained in cluster 4.
- The qualities related to tailoring and adjustment of the ML system to different environments (e.g., flexibility, adaptability) are grouped in cluster 5.
- Concerns related to privacy and security are grouped together in cluster 6.
- The NFRs that do not share similar properties are grouped in cluster 7.

We discuss these results in Sec. 5.

4.2 Estimated Number of Publications (RQ2)

We used the search strings described in Sec. 3.3 to identify an upper limit on the number of relevant publications for each NFR. The number of identified publications is presented the first column of Table 2. We found the most results for performance, accuracy, and efficiency; while, repeatability, testability, and justifiability yielded the fewest results. We found no research papers in Scopus for retrainability, potentially indicating that this term is not common.

We can examine these results from the perspective of the clusters presented previously. We can sum the total number of search results for each cluster, finding that cluster 4 (performance, ...) has 140695 results, cluster 1 (accuracy, ...) 105805, cluster 6 (security, ...) 33343, cluster 7 ("other" NFRs) 19207, cluster 5 (adaptability, ...) 6872, cluster 3 (bias, ...) has 5538 and cluster 2 (explainability, ...) has 3978 results. We discuss these results in Sec. 5.

Although we believe these results are a useful staring point, we refine our estimation for a subset of NFRs from an "upper limit" to an estimation of how many publications are actually relevant. When selecting NFRs for a more detailed estimation, we focused on NFRs that are less researched (but still potentially important), and those in Clusters 5 and 7.

We applied the inclusion-exclusion criteria and ICR process described in Sec. 3.3.3 for a sample of fifty papers of each selected NFR. We present the number of publications found to be relevant for each, along with the inter-coder reliability in columns 4-5 of Table 2. In some cases, we also conducted a second sample of an additional 50 publications. In such cases, the number of relvant publications and agreement on the second sample are listed in columns 6-7 of Table 2.

We can evaluate the strength of our agreement as follows: < 0.0 is considered as poor agreement, 0.00-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1.00 as almost perfect [20]. We attained a substantial rage of scores in terms of ICR for the NFRs. One result (ethics) was poor, with our

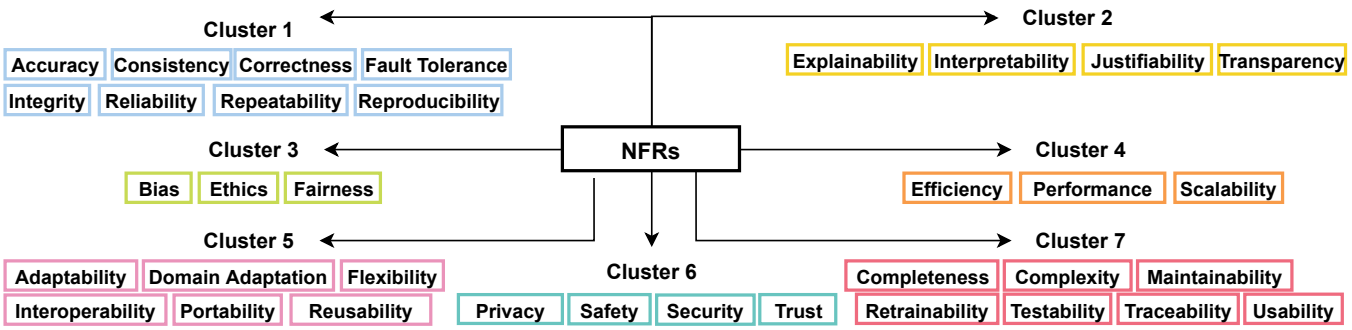

Figure 1: NFRs divided into clusters, based on shared characteristics.

Table 2: NFRs with number of search results, kappa values (agreement on sample), and final paper volume estimation for select NFRs. We only examined a second sample in cases where we wanted to see if agreement would improve.

NFR	Cluster	Search Results	Relevant Papers (Sample 1)	Fleiss' kappa (Sample 1)	Relevant Papers (Sample 2)	Fleiss' kappa (Sample 2)	Estimated Num. Relevant Pubs.
Performance	4	114853					
Accuracy	1	92669					
Efficiency	4	22247					
Security	6	19142					
Complexity	7	16997					
Privacy	6	6388					
Safety	6	5848					
Reliability	1	5620					
Bias	3	4118					
Scalability	4	3595					
Consistency	1	2936					
Flexibility	5	2764	23 (46%)	0.54			1271
Interpretability	2	2418					
Trust	6	1965					
Reproducibility	1	1796					
Domain Adaptation	5	1732	47 (94%)	0.63			1628
Usability	7	1270	21 (42%)	0.50	29 (58%)	0.44	635
Adaptability	5	1177	34 (68%)	0.50			800
Fairness	3	1089	45 (90%)	0.41			980
Correctness	1	1045	16 (32%)	0.53			334
Integrity	1	1015					
Transparency	2	851	44 (88%)	0.70			749
Explainability	2	706	44 (88%)	0.22			621
Fault Tolerance	1	553	26 (52%)	0.68			288
Interoperability	5	532	9 (18%)	0.45			96
Completeness	7	372	23 (46%)	0.40	25 (50%)	0.58	179
Portability	5	346	21 (42%)	0.45			145
Ethics	3	331	31 (62%)	-0.03			205
Reusability	5	321	24 (48%)	0.55			154
Maintainability	7	277	6 (12%)	0.30	9 (18%)	0.72	42
Traceability	7	214	4 (8%)	0.61	6 (12%)	0.61	21
Repeatability	1	171	17 (34%)	0.44			58
Testability	7	77	4 (8%)	0.54	2 (4%)	1.00	5
Justifiability	2	3	0 (0%)	1.00			0
Retrainability	7	0					0

coding being worse than random. However, we attained fair results for three other NFRs, and moderate or better for the remaining 15.

Focusing on five of the NFRs in cluster 7 as a particular example, we can examine our change in agreement after discussion. After the first sample of 50 papers, the ICR scores were fair for maintainability and completeness, and moderate for usability, completeness, maintainability, traceability, and testability. After a discussion among all three authors about our perception and interpretation of the NFR definitions and the inclusion and exclusion criteria, the ICR for the second sample generally improved and ranged between moderate (e.g., completeness, maintainability, traceability) to perfect (e.g., testability). We note, however, that our ICR score for usability actually decreased in the second round. To some extent, these score also depend on the percentage of relevant publications. The less often papers are relevant (e.g., testability), the easier it is to gain high agreement.

We estimated the total number of relevant papers for the selected NFRs according to the procedure described in Sec. 3.3. The final estimation is shown in the final column in Table. 2. We discuss these result in Sec. 5.

4.3 NFR Scoping Over System Elements (RQ3)

NFRs can be defined over different granular levels of the system. ML is a small part of a bigger system, and clear definition of NFRs in a ML system context,requires understanding which specific elements of an ML system that an NFR is applicable to. As a starting point for building this understanding, we believe that NFRs can be defined over the following parts of an ML system:

- **Training Data:** The data used by the ML algorithm as the basis for making decisions.
- **ML Algorithm:** The algorithm that performs the learning task. This includes algorithms that operate on training data, as well as those that perform learning tasks based on feedback, such as reinforcement learning agents. We also consider the specific implementation of the algorithm here.
- **ML Model**[2]**:** The core artifact built by the algorithm for use in making decisions. For example, the algorithm may use the training data to build a model that makes decisions in new situations based on learned connections between data items.
- **Results:** The resulting decisions or behaviors made as a result of applying the model.
- **Whole System:** The ML system as a whole.

These parts are illustrated in Fig. 2. It is possible that more elements may be applicable in the future, e.g., NFRs over features of a data set or over specific types of functionality operating on the results of ML, but we start with this initial list of system elements to understand the scope of the selected NFRs.

Our overall determination of which system elements a particular NFR can be defined over is presented in Table. 3. While determining over which elements an NFR can be defined over, we used the process described in Sec. 3.4.

[2]This notion also encompasses the policy learned by an agent in reinforcement learning, or other rules "learned" by the algorithm in other techniques.

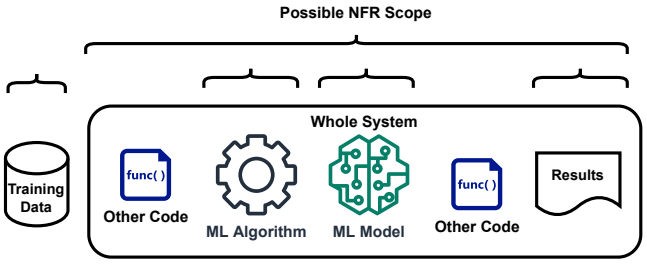

Figure 2: Possible scope for NFRs over system elements.

Table 3: System elements that NFRs can be defined over

		System Element the NFR Can be Defined Over				
NFR	Cluster	Train. Data	Algo.	Model	Results	Whole System
Completeness	1	✔	✖	✔	✖	✔
Correctness	1	✔	✔	✔	✔	✔
Fault Tolerance	1	✖	✔	✔	✖	✔
Integrity	1	✔	✔	✔	✔	✔
Repeatability	1	✖	✖	✖	✔	✔
Explainability	2	✖	✔	✔	✔	✔
Transparency	2	✖	✔	✔	✔	✔
Ethics	3	✔	✔	✔	✔	✔
Fairness	3	✔	✔	✔	✔	✔
Adaptability	5	✔	✔	✔	✔	✔
Domain Adaptation	5	✔	✔	✔	✔	✔
Flexibility	5	✖	✔	✔	✖	✔
Interoperability	5	✖	✔	✔	✖	✔
Portability	5	✔	✔	✔	✖	✔
Reusability	5	✔	✔	✔	✖	✔
Maintainability	7	✔	✔	✔	✖	✔
Testability	7	✖	✔	✔	✔	✔
Traceability	7	✔	✔	✔	✔	✔
Usability	7	✖	✔	✔	✔	✔

To illustrate our determinations, we select a number of examples. For example, we determined that the NFR flexibility can be defined over the ML algorithm, the ML model, and the whole system. However, we believe it is not applicable to the training data and the results. Consider a definition of flexibility by Ladiges et al. [19], "*flexibility is an indicator for the ability of a system to react to changing demands or conditions*". We can adapt this definition to different parts of the ML system[3]:

- Flexibility of an ML algorithm: *the ability of an algorithm to react to changing demands and conditions, without significant re-implementation.*
- Flexibility of an ML model: *the ability of a model to react to changing inputs and contexts in a useful way, without retraining.*
- Flexibility an ML system: we could keep the initial definition or make it more specific, e.g., *the ability of a ML system to react to changing demands or conditions without extensive re-implementation or re-training.*

On the other hand, we struggle to define flexibility over training data. It makes sense to think of the reusability of training data, e.g., to train ML systems for different context and purposes with some of the same data, but what does it mean for data itself to be flexible?

[3]We note that these definitions may have significant overlap with definitions for NFRs such as adaptability, resuability, or portability, which is precisely why these NFRs are placed in the same cluster—cluster 5, in this case.

Similarly, results can be reusable, but it is not clear how they can be flexible. We opt to omit these definitions from our consideration.

Similarly, the NFR usability can be defined over the ML algorithm, the ML model, the results, and the whole system; but may not be applicable over the training data. If we take the simple definition of usability from Table 1, "*how effectively users can learn and use a system*", this definition makes sense over the whole system. We can also define this NFR over specific ML elements:

- Usability of an ML algorithm: *how effectively users can learn and use an algorithm to train an ML model as part of a system.*
- Usability of an ML model: *how effectively users learn to use an ML model at run-time in order to get results.*
- Usability of ML results: *how effectively users can understand and apply ML results for some practical purpose.*

However, we struggle to create a definition for the usability of the training data. Does a user learn data? Although a user uses data, is some data more usable than others, or is that more a matter of data quality and data appropriateness? Note that this depends on what is meant by usability. When processing the abstract and titles for usability, we noted that many authors used usability as more of a binary term meaning applicability—e.g., you can use data to train a model, therefore this indicates some level of data usability. We disagree with this use of usability, and claim that usability is more appropriate as a user-centered qualitative concept. If we exclude general applicability, we find it hard to define usability of data.

Other combinations of system elements and NFRs can be defined similarly. We believe that these results can guide developers to identify different elements of the system where NFRs can be considered and assessed. The results also help them to identify and define a particular NFR that applies to a specific element of an ML system. We are working towards a framework for the definition of each NFR over each part applicable part of the system, including a checklist on which part of the system a particular NFR can be defined. We leave a full catalogue of definitions and checklist for future work.

5 DISCUSSION

In this section, we discuss our study findings. We organize this discussion by RQ.

RQ1: Clustering. We have created some initial clusters of the NFRs found relevant in our previous industrial study.

Although much previous work has explored the relationship between NFRs, usually NFRs are presented in terms of a hiearchy (e.g., [5]) or decomposed as part of a softgoal interdependeny graph (e.g., [7]). Although these approaches bear similarity to our results, our purposes were not to suggest a definite hierarchy, but to help us, and other researchers, sort relevant NFRs and future systematic mappings and NFRs. For example, a future SLR may focus on cluster 7 including usability, or cluster 5 focusing on adaptation. Alternatively, studies may focus on one or two select NFRs.

We intend for our suggested clusters to help shape and guide new research, with further SLRs or mapping studies focusing potentially on one cluster, or by using the clusters to group future related work. These clusters can also help practitioners to understand the similarity of NFRs and provide them guidance on which related NFRs they can consider in each cluster while developing ML systems.

Note that our estimation of NFR scope over system parts as indicated in Table 3 is a first pass based on our experiences and on covering the selected abstracts. The scope of NFRs can and likely will evolve over time as more data an experience is gathered, particularly as we and others find more examples, and especially metrics, for NFRs over different parts of an ML system.

RQ2: Estimation of Literature Coverage. We can reflect on the number of papers found via Scopus search, or predicted via our ICR results. The number of papers found for accuracy is very high as one would predict, as researchers and practitioners are focusing mainly on the accuracy of ML systems. We were surprised that no publication was found for retrainability even though practitioners in [11] mentioned retrainability is an important NFR for ML systems. Retrainability is a NFR which is exclusively ML-specific, perhaps these ideas are being discussed using terms which are less recognizable as an NFR. Similarly, we were surprised by so few search results on testability, but this may again be due to a different use of terms, perhaps researchers are using test or testing. We expected more search results on fairness as we perceive that researchers and practitioners are currently focusing on this topic. This may be due to the commonality and split of results amongst bias, fairness, and ethics. We found more papers for usability than we expected, even excluding for those papers using usability as applicability, but find it encouraging that research is focusing on these human-oriented aspects.

From the clustering point of view, we have listed the sum of raw search results per cluster. Although this is only an initial rough estimate of interest, we can see a particular interest in cluster 4, including performance. In ML terminology, we note that performance is often used to denote a form of accuracy, how correct the results are, as opposed to capturing running time or memory measures as the term is often used in typical SE. We note that the performance and accuracy clusters (4 and 1) show the most raw results, followed by the security cluster. These results are generally in line with our expectations. We find that cluster 7 containing those NFRs which are hard to cluster actually has a relatively large number of raw results, mainly due to the inclusion of complexity. We are particularly surprised that clusters 2 containing explainability and cluster 4 containing bias show relatively less raw results. It seems that even though these are perceived as hot topics in ML research, either the volume of papers is still relatively small, or this work includes terms which differ from the NFRs included as part of our search.

While going through the 50 titles and abstracts, we found that for some NFRs it was difficult to determine inclusion/exclusion, as is shown by our poor ICR results. For example, we had particularly low results for ethics and explainability, even though these are often perceived as popular topics in ML research. This popularity may have contributed to our difficulties. In other cases, we can often make a clear distinction between NFRs over parts of the ML system and using ML to improve an NFR in another system, e.g., traceability of ML solutions vs. the many papers using ML to find trace links. With topics like ethics and explaininability, the focus of most papers was on the ML systems themselves, i.e., many papers

were relevant. In these cases, future work may have to work harder on providing clear definitions or specific criteria.

RQ3: NFR Coverage Scope. When looking for trends or patterns in our preliminary determination of NFR scope, we can see that most NFRs can be defined over the whole system because the whole system is similar to the scope of NFRs over traditional systems. We see that all NFRs apply over the whole system, and almost all apply over the model, and most to the algorithm. On the other hand, fewer NFRs can apply to the training data and the results, but there is no clear pattern here. We do not find that NFRs always can be applied to the results and data together, sometimes it is one or the other or both. We intend to continue this work towards a framework to guide NFR exploration, consideration and measurement over ML systems. We hope that such exploration can lead to a deeper understanding of what makes NFRs applicable to different system parts, and why.

5.1 Threats to Validity

External Validity: In terms of publication database, we have only used Scopus, which may mean we miss relevant papers in other databases that could be important for our study. However, Scopus is a meta-database that is rich in content and relatively inclusive for computer science research, and it includes peer-reviewed publications from multiple publishers. We searched papers in Scopus up to September 2021, which means that there may now be more research papers related to our study that are missed. Future secondary studies should repeat the search process.

The search string was confined to a small set of search terms and keywords, focusing on only a subset of NFRs, but the search terms were inspired by our research questions and the findings from our previous industrial study. As an example, we could have also searched for NFR stems like "interoper" for interoperability. However, it would be difficult to find equivalent stems for all NFRs (e.g., security) and may have led to an unmanageable increase in the volume of papers without a significant increase in relevant results. Our goal is not to make a conclusive statement on the number of publications that exist for each NFR, but to gain an approximate idea of the level of past interest in each. A sample of publications is sufficient for such purposes.

Internal Validity: There is potential bias in determining paper inclusion. To mitigate this risk, we defined shared inclusion and exclusion criteria, each of the three authors went through each title and abstract separately, and in cases where we disagreed, we read the title and the abstract of the paper together again, to make the final decision. Our ICR results are often in a good range, and repeating these results on a second sample for selected NFRs yielded consistent or better ICR scores for all but one NFR.

The clusters of NFRs we created may be subjective to our experiences and opinions. One may argue the NFRs could be arranged differently, but we believe our clusters are a good starting point to help organize and direct future research. Further work may add to or adjust the clusters as new evidence is found.

Our consideration of the scope of NFR definitions may also be subjective and hard to justify. We made these judgements in a meeting with the three paper authors, discussing difficult cases. We have tried to justify our selection for a sample of NFRs. Future work can adjust our initial scoping decisions when more evidence or examples are found.

6 CONCLUSIONS

In this work, we aimed at understanding and exploring definitions, scope, and the extent of existing research on NFRs for ML systems, as we believe that both the research community and industry lack knowledge on NFRs for ML systems compared to understanding in traditional systems. The results show that researchers have focused on many NFRs for ML systems, but the amount of attention directed to each NFR differs drastically. Some NFRs received more attention and were explored more (e.g., peformance, accuracy, efficiency) compared to other NFRs (e.g., maintainability, traceability). Although such differences were expected, it is useful estimate interest with concrete numbers.

We created six clusters of NFRs based on the similarity of characteristics and meaning of NFRs, and one cluster of NFRs which does not share similar properties, with the objective of helping researchers to focus on a particular cluster for their future systematic review studies. These clusters will also help practitioners to understand the similarity of NFRs and provide them a direction on which NFRs they need to consider while developing ML systems.

We defined NFRs over different granular levels of the ML systems based on the meaning and purpose of those NFRs. This can help practitioners to understand on which part of the ML system a particular NFR can be considered while developing ML systems. Our future work includes a comprehensive mapping study to identify the current state-of-the-art on selected NFRs for ML systems research, and a framework to guide consideration of NFRs over different elements of ML systems.

REFERENCES

[1] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. PMLR, 149–159.

[2] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 8–13.

[3] Barry W Boehm, John R Brown, and Mlity Lipow. 1976. Quantitative evaluation of software quality. In *Proceedings of the 2nd international conference on Software engineering*. 592–605.

[4] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 1175–1191.

[5] Joseph P Cavano and James A McCall. 1978. A framework for the measurement of software quality. In *Proceedings of the software quality assurance workshop on Functional and performance issues*. 133–139.

[6] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 28, 3 (2019), 231–237.

[7] Lawrence Chung, Brian A Nixon, Eric Yu, and John Mylopoulos. 2012. *Non-functional requirements in software engineering*. Vol. 5. Springer Science & Business Media.

[8] Joerg Doerr, Daniel Kerkow, Tom Koenig, Thomas Olsson, and Takeshi Suzuki. 2005. Non-functional requirements in industry-three case studies adopting an experience-based NFR method. In *13th IEEE International Conference on Requirements Engineering (RE'05)*. IEEE, 373–382.

[9] Jonas Eckhardt, Andreas Vogelsang, and Daniel Méndez Fernández. 2016. Are" non-functional" requirements really non-functional? an investigation of non-functional requirements in practice. In *Proceedings of the 38th International Conference on Software Engineering*. 832–842.

[10] Martin Glinz. 2007. On non-functional requirements. In *15th IEEE International Requirements Engineering Conference (RE 2007)*. IEEE, 21–26.

[11] Khan Mohammad Habibullah and Jennifer Horkoff. 2021. Non-functional Requirements for Machine Learning: Understanding Current Use and Challenges

in Industry. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*. IEEE, 13–23.

[12] Jennifer Horkoff. 2019. Non-functional requirements for machine learning: Challenges and new directions. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*. IEEE, 386–391.

[13] Fuyuki Ishikawa and Nobukazu Yoshioka. 2019. How do engineers perceive difficulties in engineering of machine-learning systems?-questionnaire survey. In *2019 IEEE/ACM Joint 7th International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP)*. IEEE, 2–9.

[14] Aleksander Jarzębowicz and Paweł Weichbroth. 2021. A Systematic Literature Review on Implementing Non-functional Requirements in Agile Software Development: Issues and Facilitating Practices. In *Lean and Agile Software Development*, Adam Przybyłek, Jakub Miler, Alexander Poth, and Andreas Riel (Eds.). Springer International Publishing, Cham, 91–110.

[15] Toshihiro Kamishima, Shotaro Akakamishima2011fairnessho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

[16] Staffs Keele et al. 2007. *Guidelines for performing systematic literature reviews in software engineering*. Technical Report. Citeseer.

[17] Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. (2007).

[18] Zijad Kurtanović and Walid Maalej. 2017. Automatically classifying functional and non-functional requirements using supervised machine learning. In *2017 IEEE 25th International Requirements Engineering Conference (RE)*. Ieee, 490–495.

[19] Jan Ladiges, Alexander Fay, Christopher Haubeck, and Winfried Lamersdorf. 2013. Operationalized definitions of non-functional requirements on automated production facilities to measure evolution effects with an automation system. In *2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA)*. IEEE, 1–6.

[20] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[21] Payman Mohassel and Yupeng Zhang. 2017. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 19–38.

[22] Bashar Nuseibeh and Steve Easterbrook. 2000. Requirements engineering: a roadmap. In *Proceedings of the Conference on the Future of Software Engineering*. 35–46.

[23] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[24] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015), 2503–2511.

[25] Andreas Vogelsang and Markus Borg. 2019. Requirements engineering for machine learning: Perspectives from data scientists. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*. IEEE, 245–251.