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)

How do Engineers Perceive Difficulties in Engineering of Machine-Learning Systems? - Questionnaire Survey

Fuyuki Ishikawa
National Institute of Informatics
Tokyo, Japan
f-ishikawa@nii.ac.jp

Nobukazu Yoshioka
National Institute of Informatics
Tokyo, Japan
nobukazu@nii.ac.jp

***Abstract*—There is increasing interest in machine learning (ML) techniques and their applications in recent years. Although there has been intensive support by frameworks and libraries for the implementation of ML-based systems, investigation into engineering disciplines and methods is still at the early phase. The most pressing issue in this field is identifying the essential challenges for the software engineering research community as engineering of ML-based systems requires novel approaches due to the essentially different nature of ML-based systems. In this paper, we analyze the results of a questionnaire administered to 278 people who have worked on ML-based systems in practice, clarify the essential difficulties and their causes as perceived by practitioners, and suggest potential research directions.**

***Index Terms*—software engineering, machine learning, artificial intelligence, questionnaire survey**

I. Introduction

Practical and industrial applications of machine learning (ML), or, in the broader sense, artificial intelligence (AI), have been actively investigated. More and more engineers are being faced with increasing demands for the high quality of such applications as well as greater efficiency of the development and operation processes. Obviously, engineering disciplines and techniques are crucial to support the widespread use of ML or AI systems in society.

One of the most pressing issues in this field is identifying the essential challenges facing the software engineering research community (SE). In other words, are there critical differences that require us to invent novel disciplines? The most significant point is that, when we use ML techniques, the behavior of a software component, e.g., a neural network or a decision tree, is inductively derived from training data. This is different from the standard situation in SE, where the behavior of the component is governed by rules that human engineers deductively design. It is not a great stretch to suggest that we are dealing with a different paradigm of software construction. At the very least, in engineering we simply have a different target - specifically, training data - due to the difference of the construction process. Moreover, unique essential difficulties have been discussed, such as the unexplainability of individual outputs [1], weakness against slight changes of inputs (adversarial examples) [2], and the changing anything changes everything (CACE) principle (an example of unique technical debts) [3]

Research efforts to tackle such difficulties have been emerging over the past few years but are still limited. A few simple guidelines have been put in place by leading companies [3]–[5]. One recognizable trend is the emergence of research on the testing and verification of components and systems constructed by ML techniques [6]–[11]. However, the scope of these studies is limited compared with the wider application area of SE. It is essential to clarify what type of difficulties engineers are facing so that effective research directions can be identified.

In this paper, we analyze the results of a questionnaire that covers the engineering of ML techniques. This questionnaire targeted 278 individuals who are working or have worked on ML or ML applications in practice (not for learning but for products or services used by other people) in Japan. Most of the participants are engineers who have extensive experience with software development and have recently started using ML techniques. We summarize the insights obtained from our analysis of the questionnaire results and discuss its implications for the SE research community.

In Section II of this paper, we give an overview of the research on the engineering of ML-based systems to date. Then we explain the questionnaire methodology and briefly provide the simple statistical results in Sections III and IV. In Section VI, we go over the results in depth for the core part: specific goals and difficulties of ML-based systems. We discuss insights obtained from the results in Section VII and conclude with a brief summary and mention of future work in Section VIII.

II. Background and Related Work

In this paper, we use the term *ML components* to denote components obtained by ML techniques, e.g., decision trees obtained by the Random Forest method[1]. We call systems that include ML components *ML-based systems*, and these are assumed to be affected by the nature of ML techniques.

[1] We do not call this a “model”, even though it is called this in the ML community, so as to avoid confusion.

2575-4793/19/$31.00 ©2019 IEEE
DOI 10.1109/CESSER-IP.2019.00009

A. Insights from ML Community

The majority of research in the ML community has targeted algorithmic aspects for performance (typically accuracy and the like), as well as libraries and platforms to facilitate the implementation. Given the increasing use of ML in society, quality and dependability issues are being actively discussed.

One of the most notable directions in the ML community is Explainable AI (XAI) [1]. This direction is to tackle difficulties caused by the nature of ML, especially deep learning, which is black-box and lacks any explanation about why the output is obtained. For example, the work in [12] generates an area of images that explains the results of image classification.

A well-known problem of ML, though discovered recently, is the existence of adversarial examples [2]. Slight changes in the input, even if they are not recognizable by humans, can radically change the output with the most sophisticated image classifiers. This suggests the need for specific techniques for testing and verification, as discussed in Section II-B.

Even if unexpected behavior were not an issue, achieving high accuracy is difficult in general and engineers may be faced with unexpected difficulties or limitations. A recent example is the problem of socially inappropriate tagging in Google Photos, which could not be essentially resolved[2]. This point highlights the uncertainty or unpredictability [that is often present?] in the development process.

Leading companies such as Google have published problem statements and guidelines based on their experience [3]–[5]. For example, the need for runtime monitoring is emphasized given the strong dependency on external factors such as input distribution. Although the insights given in these studies are undoubtedly supported by rich experience, scientific or empirical analysis has not yet been provided.

There have been studies on challenges regarding activities or concerns specific to the data-centric nature of ML, e.g., data management [13] and security [14]. In this paper, we focus on new challenges in SE activities.

B. Testing and Verification of ML

A notable movement is the emergence of testing and verification methods for ML components and ML-based systems.

Coverage criteria have been used as core metrics to capture the effectiveness of tests in terms of exhaustiveness or diversity. Studies in [6] and [7] used a unique kind of coverage criteria, neuron coverage, for neural networks. The neuron coverage was utilized in search-based testing to find diverse problematic inputs, e.g., inputs that lead to different behaviors between the component under testing and one for comparison.

Formal verification of ML components has been investigated to check the robustness or non-existence of adversarial examples [8], [9]. Various techniques have been used, including SMT solvers, abstract interpretation, and stochastic games. These studies succeeded in detecting unexpected results, e.g., adding a one-pixel change in an image changed the result of road signal recognition.

The work in [10] examined system-level testing. Specifically, it extracted and focused on the possibility that failures of the ML component lead to failures of the system. This is in contrast to most of the investigation efforts in the ML community, where adversarial examples are typically discussed for the classification of arbitrary images at the ML component level. This direction is significant in terms of linking requirements and environmental assumptions to the quality of the ML component.

Another approach is metamorphic testing [15]. It is difficult or practically impossible to define the expected output for an arbitrary input to an ML component or an ML-based system (i.e., it is a non-testable program [16]). In metamorphic testing, we check an expectation that "changing the input in this way should change the output in that way" (metamorphic relation). This approach has been investigated for ML systems, e.g., [11].

III. Questionnaire Methodology

We base our analysis and discussion on a questionnaire created by JSSST-MLSE, an academic society in Japan[3]. JSSST-MLSE, established in April 2018, aims to provide venues for researchers and practitioners to share and discuss research studies and practices.

The objective of the questionnaire was to grasp the present state of ML applications from the viewpoint of SE activities and to clarify how the characteristics of ML are perceived. The questionnaire covered the following aspects.

- Experience with SE activities
- Experience with ML techniques
- Past projects that used ML and quality attributes that were considered significant in the projects
- Perception of difficulties in the engineering of ML-based systems
- Characteristics of ML that lead to the difficulties
- Quality attributes that will be significant in future projects

The questionnaire was disseminated publicly on the Web. We explained that the target is people who have used ML in their work. The request to answer the questionnaire was sent out via mailing lists and social networks. The target mailing lists included those of academic societies (including MLSE) pertaining to AI/ML or SE, as well as practical societies such as the Japan Deep Learning Association[4]. The account of JSSST-MLSE was used in the dissemination to social networks. A total of 279 answers were collected in the target period of about one month. The results are available online[5] and we use this data for our analysis and discussion in this paper.

[2] https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people

[3] The Special Interest Group on Machine Learning Systems Engineering, Japan Society for Software Science and Technology. https://sites.google.com/view/sig-mlse/ (in Japanese).

[4] Providing qualification on deep learning techniques, http://www.jdla.org/en/

[5] The statistics are available at the aforementioned Web site of JSSST-MLSE (in Japanese) and detailed data including free descriptions are available by querying to JSSST-MLSE for the purpose of research.

TABLE I
EXPERIENCE OF PARTICIPANTS

ML Dev	L1 (-1Y)	L2 (1-3Y)	L3 (3-10Y)	L4 (10Y-)	Total
L1 (-1Y)	21	5	1	0	27
L2 (2-3Y)	15	19	6	1	41
L3 (4-5Y)	9	9	12	3	33
L4 (6-10Y)	17	20	12	2	51
L5 (10Y-)	36	52	18	15	121
Total	98	105	49	21	273

TABLE II
SECTOR OF PARTICIPANTS

Sector	Rate [%]
Manufacturing	29.8
Information Processing & Provisioning Services	21.9
Software	21.5
Internet Services	6.1
Education (University, etc.)	5.3
Finance & Insurance	3.2
Services	2.8
Others (Public, Wholesale & Retail, etc.)	9.4

IV. BASE STATISTICS

First we describe the profiles of the participants and what type of ML applications they have worked on.

A. Participants

Table I shows the length of experience the participants have for "activities for software development & application" and ML. The participants chose from among five levels for development (Dev) and four levels for ML. We can see that the majority of the participants are engineers who have extensive experience with software development: the total of Dev L4–5 is 172. Many of them newly started to work on ML (the total of Dev L4–5 & ML L1–2 is 125). In contrast, few people are involved who have long experience with ML but not with SE: the total of Dev L1–2 & ML 3–4 is 8. This fact shows that very few scientific ML researchers who typically work on theories, algorithms, and proof-of-concept applications are included, rather than engineering for products and services.

Table II shows the sectors of the participants. The rate of participants from the Education section, probably university researchers, is not so large. We thus expect that most of the participants are from the industry. The high rate of those from Manufacturing is notable and relevant to the country, Japan, where the questionnaire was administered.

From these results, we conclude that the participants constitute a sufficient range of types to obtain insights into the practical engineering of ML-based systems.

B. ML Applications

Figure 1 shows the application domains of ML projects that participants have joined (one participant may have joined multiple projects). The outstanding one is again the Manufacturing domain, which features 85 participants, or about one third of all participants. The experiences of the participants cover various other domains as well.

V. QUALITY ATTRIBUTES

Figure 2 shows the results of questions about the significance of different quality attributes. Specifically, we asked 1) which attributes were thought to be significant in past projects and 2) which they felt would be significant in the future. We expected these two questions reveal possible immaturity of past projects and foresight for improvement. Each participant could choose multiple attributes, and the vertical axis in the figure represents how many participants chose the attribute. The results are sorted by the rates for the first question.

"Understandability & Explainability of Outputs" were considered significant by many of the participants (71% for the past and 64.7% for future). This fact demonstrates the active trend in XAI in the ML research community and is supported by strong demand. The four attributes on the right-hand side of the figure are those that were left behind in past projects (Long-term Maintenance and Adaptation to Changes, Safety, Security, and Privacy). This result suggests some kinds of immaturity. There might be limited application areas where such quality attributes are less significant. Another possibility is that many engineers were struggling to construct ML applications that function well, and they could not expend much effort on preparing for future changes or the mitigation of various risks. Nevertheless, the significance of these attributes is recognized to some extent, as indicated by the increase of rates in the future expectation.

VI. DIFFICULTIES OF ENGINEERING OF ML-BASED SYSTEMS

A. Difficulty Levels

As the core of the questionnaire, we asked the participants about the level of difficulty in activities related to the engineering of ML-based systems. Specifically, we had them choose one of the following four levels of difficulty for each activity.

- L4: Existing Approaches Not Applicable Anymore
- L3: Same Approaches Applicable but Methods/Tools Immature
- L2: Dedicated Methods/Tools Available
- L1: Existing Methods/Tools Applicable

Figure 3 shows the answers to the questions about difficulty level. The results are sorted by the rate of those who chose L4. In general, difficulty levels are felt to be high. L3 or L4 was selected by more than 60% of participants for all items and by around 80% for half of them. L4 was selected by more than 40% of participants for the top two activities: "Decision Making with Customers" and "Testing & Quality Evaluation / Assurance". These points strongly demonstrate the immaturity of engineering for ML-based systems.

Some activities were thought less difficult. For example, "Architecture Design" is considered the least difficult. This is probably because new design patterns should be mined, collected, and disseminated for new kinds of systems. However, the way to do so is not new and we can expect new design patterns are obtained through accumulation of experience.

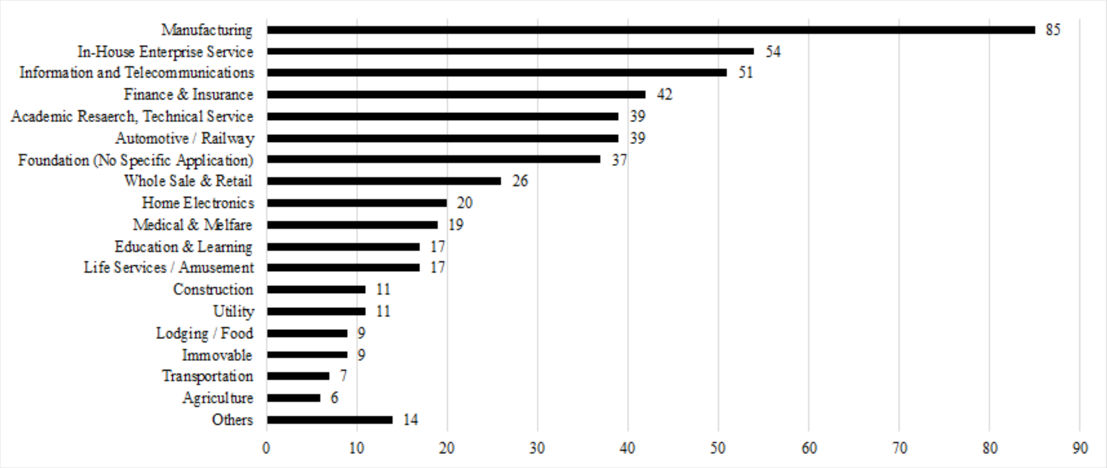

Fig. 1. ML Application Domains

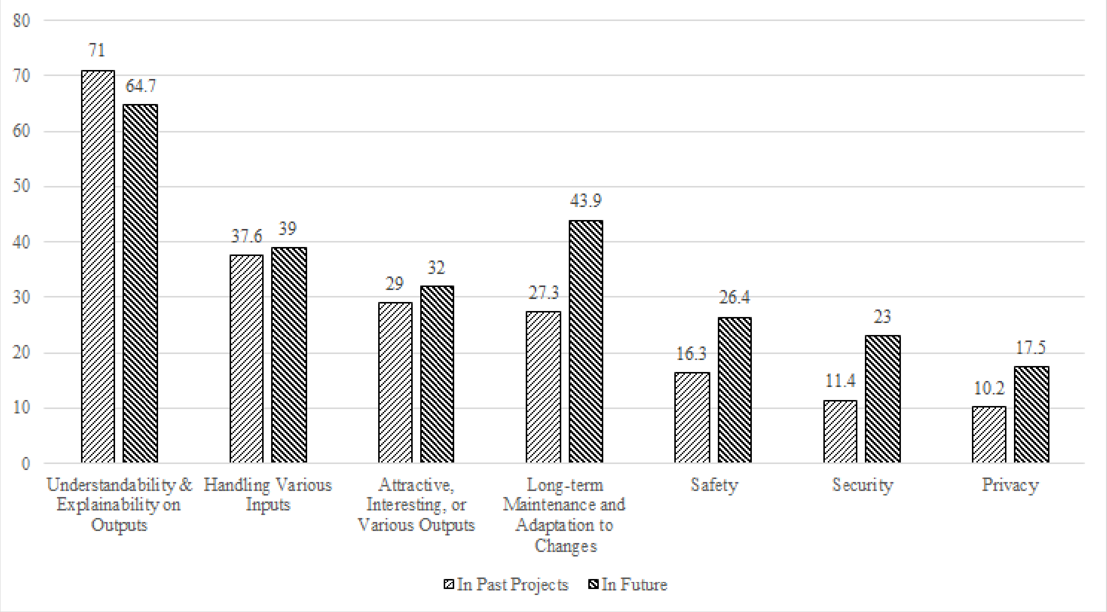

Fig. 2. Quality Attributes Considered Significant in Past and in Future

In the later analysis in Section VI-C, we will especially focus on the top two activities, which also received many comments. In the remainder, we will abbreviate these two activities as *decision making* and *testing & QA*, respectively.

B. Sources of Difficulty

Figure 4 shows the answers to the question about what the most significant causes of the difficulties are (at most, the top 3 for each participant). The item names are abbreviated in the figure and detailed descriptions of the choices are shown in Table III.

The first four items represent the specific characteristics of ML and are thought to be essential causes of the difficulties. The fifth item, openness, is shown to be less significant. This is probably because some applications are operated in closed environments (e.g., a factory) and assumptions are somewhat clear. This cause has been around since before the emergence of ML, e.g., in arguments on car safety. The last item, expertise, is shown to be the least significant. This result may suggest that engineers do not hesitate to learn new theoretical backgrounds, or, in contrast, that they do not respect the necessity of learning the theoretical backgrounds.

In this paper, we will not get into the less significant points, such as the last point about the perception of expertise. We focus on the most significant points indicated by the participants, as described in the following analysis and discussion.

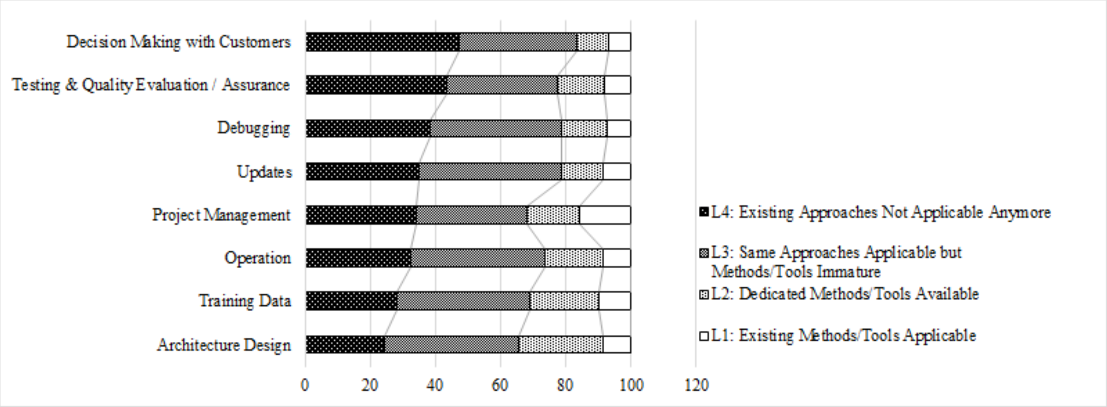

Fig. 3. Perception of Difficulties

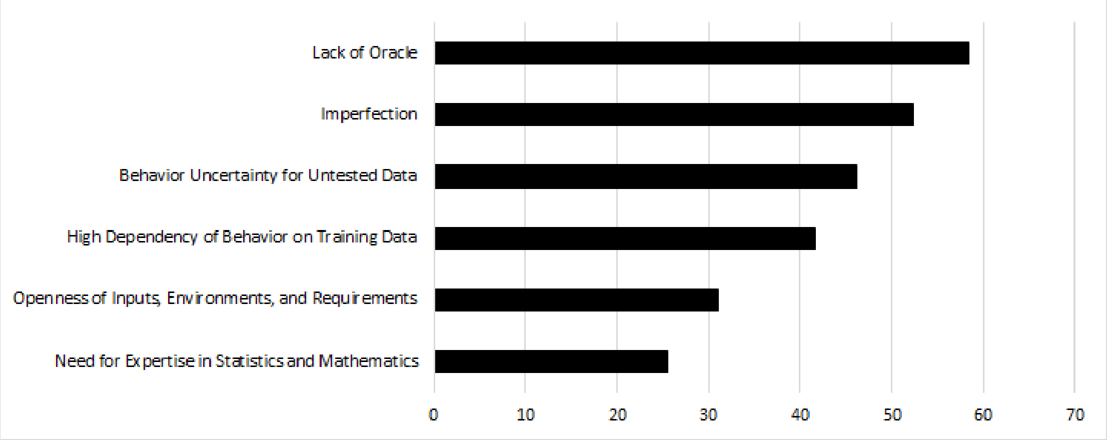

Fig. 4. Causes of Difficulties

TABLE III
DETAIL OF CHOICES FOR CAUSES OF DIFFICULTIES

Lack of Oracle	It is difficult or impossible to clearly define the correctness criteria for system outputs or right outputs for each individual input.
Imperfection	It is intrinsically impossible to make adequate outputs for any of various inputs (i.e., 100% accuracy).
Behavior Uncertainty for Untested Data	The uncertainty is high about how the system behaves in response to untested input data, such as radical change of the behavior by slight change in the input (adversarial examples).
High Dependency of Behavior on Training Data	The system behavior highly depends on the training data.
Openness of Inputs, Environments, and Requirements	It is difficult to have comprehensive consideration because there are enormous numbers of inputs to the system, target environments, and implicit user requirements.
Need for Expertise in Statistics and Mathematics	Expertise in statistics and mathematics is required.

C. Perceptions of Difficulties and Causes

The questionnaire aimed at grasping the overall trend of perceptions of difficulties and causes. It did not include a deep investigation into each item, e.g., "Why do you think *decision making* is difficult" or "Why do you think *lack of oracle* essentially introduces difficulties," not to put too much burden on the questionnaire participants. Instead, we included a free-description question: "What are the reasons behind your previous answers on the difficulties and causes?". In this way, hints are given regarding the point(s) that each participant felt was the most significant. We could obtain insights primarily regarding the activities of the top two difficulty levels, i.e., *decision making* and *testing & QA*, as discussed below. There were 127 answers on this free description, which represents almost half of the participants.

We first exclude general comments that apply to almost all activities. There were 10 comments that just stated immaturity of both customers and engineering teams, which led to lack of practices, a lot of trials-and-errors, and later recognition of necessary aspects. Participants who gave these comments

TABLE IV
COMMENTS REGARDING DECISION MAKING

Comment Group	#Participants
Gaps with Customers	27
- Insufficient Understanding of Customers	(20)
– Too Much Expectation	(9)
Imperfection and Accuracy Assurance	7
Uncertainty in Input-Output Relationships	3
Need for Continuous Engineering	3
Impossibility of Prior Guarantee	3
Explanation on Failure	2

chose L3 or L4 for almost all the activities, thus feeling (superficially or essentially) the comprehensive paradigm shift. Similarly, we skip discussion on 2 comments on the cultural inflexibility and work without precedents.

1) Comments on Decision Makings: *Decision making* was felt to be the most difficult, with the maximum number of participants who chose L4 (Section VI-A, Fig. 3). For further analysis on this point, we extracted free-description texts that mention phrases related to customers and end users, including indirect phrases such as "acceptance inspection". We then grouped the comments in a bottom-up manner. In cases where one participant gave multiple comments, we separated them. The resulting groups, with more than one participant, are shown in Table IV. Each of these are detailed and discussed below.

a) Insufficient Understanding of Customers: Many comments mentioned a gap between the engineering team and customers, and most of them mentioned insufficient understanding of customers. This does not necessarily mean that the customers should be blamed, as the essential cause can be an inevitable gap due to differences in goals, skills, and experience. In fact, some comments alluded to such gaps without any wording to blame the customers. There were comments about customers with unrealistic expectations, such as those who believe in a "perfect AI" that makes no mistakes. At the very least, we can see quite clearly that the gap between the engineering team and the customers is currently most significant in *decision making*, which is making many engineers stressed.

b) Too Much Expectation: Among the comments on insufficient understanding of customers, almost half of them mention too much expectation regarding the functionality, the achievable accuracy, or easy start (e.g., work with little data). The coming transition on the hype cycle may eventually resolve this issue.

c) Imperfection and Accuracy Assurance: Among the various characteristics of ML, this point was often mentioned as a difficulty. Comments mentioned that customers tend to require 100% accuracy as the standard in their mind or to require a certain assurance of accuracy, even if they understand it is impossible in theory. Essential research topics such as how to convince customers about the usefulness even with the imperfection and how to mitigate the risks of the imperfection are suggested.

TABLE V
COMMENTS REGARDING TESTING & QA

Non-Testable and Non-Deterministic	5
Unclear coverage	3
Assurance on Safety and Stability	3
No Specification as Assurance Criteria	3

d) Uncertainty in Input-Output Relationships: A few comments stated this point as a difficulty. No specific comments about why or in what sense were obtained. It seems that this point in ML-based systems, in contrast to more conventional software systems, is stirring up more confusion or anxiety in customers.

e) Need for Continuous Engineering: A few comments stated that there is a challenge in convincing the customers to keep paying continuously, as continuous improvement is desirable for imperfect ML-based systems and the systems can be easily invalidated by trend changes. It seems that engineers who perceived this concern as a core one, probably after obtaining acceptance of an initial development and release, are advanced and limited. In future, the significance of this concern is expected to grow.

f) Impossibility of Prior Guarantee: few comments stated that essential differences include engineers not being able to make any prior guarantee about the accuracy, the development time and cost, and the cost-effectiveness. One notable comment highlighted a challenge in how to remove anxiety due to this uncertainty, which seems very worrisome for customers.

g) Explanation on Failure: A few comments mentioned that it is difficult to explain failure in making a certain output. This was said to be problematic when the output from the system is counter-intuitive for human experts and when the planning for the fix is done. This point suggests that the black-box nature of ML introduces difficulty in decision making or in communication with the customers.

h) Others: There were a variety of other comments. Some of these, which we feel are essential and require investigation or changes to the current way of thinking, are listed below.

- It is unclear how we should decide on the data set used for accuracy evaluation, how to reach an agreement with the customers on this point, and who is responsible for the agreement.
- We can evaluate the benefits only by using it in actual business.
- There is no well-known agreement about how much we should test.

2) Comments on Testing & QA: We extracted groups of comments for *testing & QA* in the same way as for *decision making*. There was no strong trend among the comments for *testing & QA*, and relevant comments were diverse. The resulting groups, with more than one participant, are shown in Table V. Each of these are detailed and discussed below.

a) Non-Testable and Non-Deterministic: There were several comments that identified this point as the source of essential changes. Most of these comments were given in connection to decision making: the way of assurance will be different from conventional ones, as the requirements definition and agreement will be different. However, we could not obtain further insights about concrete difficulties that are being faced, as most comments simply restated the causes of the difficulties.

b) Unclear coverage: A few comments mentioned the aspect of "coverage." The comments highlighted the difficulty in understanding and effectively using what correspond to classical notions of coverage for requirements or implementation. This point seems open, though white-box coverage criteria have already been discussed (Section II-B).

c) Assurance on Safety and Stability: A few comments mentioned the necessity of the assurance of safety and stability for the applications they had in mind, for which principles have not been established.

d) No Specification as Assurance Criteria: There were a few comments that the way of thinking will change due to the fact that we cannot make clear specifications from which the testing & QA activities can be derived. Guidance in such situations is desired so as to avoid practices that are too ad-hoc.

e) Others: Some of the other comments about *testing & QA* that we feel are essential and require investigation or changes to the way of thinking are listed below.

- Quality management is done through preparation of training data, which is a new area.
- Offline testing without connecting to a real environment is insufficient.
- We expect the emergence of different approaches to *testing & QA* for the ML component and for the whole ML-based system.

3) Other Comments: There were comments on other activities, such as the difficulty of debugging, specifically fault localization. In general, the numbers of comments follow the perceived difficulty levels and causes (Sections VI-A and VI-B). Below, we pick up some of the other comments, which are not mere restatements of the difficulties or causes.

- General methodologies do not exist and it is necessary to discuss approaches to individual cases, which are leading to very high costs.
- What is important is fusion with the human system.
- Implementation activities are too ad-hoc and lack engineering disciplines, though this should be problematic in tackling the difficulty that the behavior radically changes depending on inputs, parameters, and random seeds.
- The speed of technical advancement is very fast compared with that for conventional techniques, which is increasing the difficulties.
- The top-down development culture, typical in so-called system integrators in Japan, is facing its own essential limitations.
- Operation plays a different role in ML-based systems, as unexpected behavior is more likely compared with conventional systems.

VII. Discussion

A. Result Summary and Suggestions

1) Current Status: Practitioners with rich experience in SE have been driven to start working with ML. A variety of applications are being investigated, such as manufacturing, which is representative in the specific context of this paper. Many of the practitioners are still at the early phase, even struggling at the POC phase.

The dominant concerns in these early explorations pertain to decision making with the customers. In the conventional setting, this activity involved requirements analysis and specification in the initial phase and an acceptance inspection in the final phase. This activity flow is not possible when working with ML-based systems due to the impossibility of prior estimation or assurance of achievable accuracy. Regardless of these facts, it seems that a non-trivial number of customers have a too idealistic or vague understanding and expectation of AI/ML, which lead to unnecessary cost, loss of chance, and disappointment. This point needs to be overcome before tackling the essential difficulties.

An immediately required action in the short term is the provision of fair and precise guidance about ML for non-technical people. Engineers have been able to grasp the nature of ML due to their concrete implementation experience, for which rich guidance and support have been provided. However, non-technical individuals are likely to be confused by the various streams of information on general Artificial Intelligence, human-beating game players, and so on. They should be guided in such a way that they can precisely distinguish relevant topics from irrelevant ones, probably focusing on statistical ML for solving business problems given the current trend. In other words, help is needed in order to clarify for these individuals what is possible and what is not.

2) Decision Making with Customers: If the customers have a better understanding of, for example, imperfection, we can move on to tackle the essential difficulties stemming from imperfection. Most of the perceived difficulties are about the uncertainty: e.g., we cannot know the value of something before we actually construct it, and we cannot be fully confident of the value because the functionality is imperfect and sometimes does not work.

This kind of difficulty can be addressed using trial-based processes that iterate the refinement of assumptions and goal settings by means of experiments. Investigation of such flexible processes is often driven by the industry and practitioners. However, support from the academic and scientific research communities is also desirable to build a solid foundation for the engineering disciplines.

3) Testing & QA: Many practitioners pointed out changes that have occurred in the way of thinking for testing and quality evaluation / assurance. This point is due to the unique nature of ML, for example, the imperfection and lack of oracle invalidated unit testing, whereas conventionally, if we found a failed test then we could say there is a bug. New disciplines as well as empirical studies are required in order to clarify the principles involved in the testing and quality evaluation / assurance of such systems.

Although there have been active research efforts focused on testing and verification, many of the studies are limited to finding adversarial examples by slight changes of the inputs. Practical concerns seem to be how much test data should be used, how to evaluate the "coverage" of the test data (e.g., against the reality), how we should judge the sufficiency of tests, and so on. Finding adversarial examples is significant but it seems that many practitioners are struggling at some point before doing it. Again, principles and empirical studies, not only individual specific testing methods, are highly demanded.

4) Machine Learning Systems Engineering: Last but not least, new concerns for practitioners include difficulties with monitoring unexpected behavior, fault localization, and maintainability for continuous engineering. Such problems are only expected to grow, as many products and services will be coming into use and the demand for quality will increase. Active research is necessary for the whole area of "Machine Learning Systems Engineering," which should be considered as a new paradigm that requires non-trivial research efforts.

B. Limitations and Treats to Validity

Readers should be aware of potential limitations in the obtained results. First, the analysis and discussion in this paper are supported by the answers to a questionnaire and thus cannot strongly support what the participants are overlooking or undervaluing. Second, the obtained insights may not apply to countries other than Japan.

The questionnaire was designed to be simple so as to reduce the burden on participants. As such, it was not possible to uniquely identify which causes were thought to lead to the difficulty of each activity type and which comments were about specific difficulties or causes (Sections VI-B and VI-C, respectively). While we endeavored to be objective and systematic in deriving the results, it is possible that our interpretations may slightly differ from the intention of the participants. We did not obtain characteristics that might affect the answers, such as applied development process or project size, either. Nevertheless, we believe that the results are sufficient for the goal of this paper, which was to provide initial insights on the engineering of ML-based systems as a whole. These results should provide both ourselves and other researchers with a good starting point for further analysis.

VIII. Concluding Remarks

We have reported and discussed how practitioners currently perceive the difficulties and their causes in the engineering of ML-based systems. Essential difficulties stemming from the specific nature of ML include imperfection, lack of oracle, and uncertainty of the implemented behavior. These causes have non-trivial effects that invalidate existing approaches, for example, in the way of decision making with customers. Although there are emerging research efforts, gaps exists between the perceived difficulties and the current scope of the research efforts.

Our future work will include expanding the survey target to outside of Japan so as to obtain more data for statistical analysis and comparison between countries. Our investigation will target the specific focuses revealed by the analysis in the present work as well as further analysis such as association analysis.

Acknowledgement: The authors are grateful to JSSST-MLSE for administering the questionnaire and for providing the results, as well as to the survey participants for their useful insights.

References

[1] D. Gunning, "Explainable artificial intelligence (XAI)," IJCAI 2016 Workshop on Deep Learning for Artificial Intelligence (DLAI), July 2016.

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, May 2015.

[3] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young, "Machine learning: The high interest credit card of technical debt," in *NIPS 2014 Workshop on Software Engineering for Machine Learning (SE4ML)*, December 2014.

[4] M. Zinkevich, "Rules for reliable machine learning: Best practices for ML engineering," NIPS 2016 Workshop on Reliable Machine Learning in the Wild, December 2017.

[5] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, "What's your ML test score? a rubric for ML production systems," NIPS 2016 Workshop on Reliable Machine Learning in the Wild, December 2017.

[6] K. Pei, Y. Cao, J. Yang, and S. Jana, "Deepxplore: Automated whitebox testing of deep learning systems," in *The 26th Symposium on Operating Systems Principles (SOSP 2017)*, October 2017, pp. 1–18.

[7] Y. Tian, K. Pei, S. Jana, and B. Ray, "DeepTest: automated testing of deep-neural-network-driven autonomous cars," in *The 40th International Conference on Software Engineering (ICSE 2018)*, May 2018, pp. 303–314.

[8] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, "Safety verification of deep neural networks," in *The 29th International Conference on Computer Aided Verification (CAV 2017)*, July 2017, pp. 3–29.

[9] M. Mirman, T. Gehr, and M. Vechev, "Differentiable abstract interpretation for provably robust neural networks," in *The 35th International Conference on Machine Learning (ICML 2018)*, vol. 80, July 2018, pp. 3578–3586.

[10] T. Dreossi, A. Donzé, and S. A. Seshia, "Compositional falsification of cyber-physical systems with machine learning components," in *The 9th NASA Formal Methods Symposium (NFM 2017)*, May 2017, pp. 357–372.

[11] A. Dwarakanath, M. Ahuja, S. Sikand, R. M. Rao, R. P. J. C. Bose, N. Dubash, and S. Podder, "Identifying implementation bugs in machine learning based image classifiers using metamorphic testing," in *The 27th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2018)*, July 2018, pp. 118–120.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, August 2016, pp. 1135–1144.

[13] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data management challenges in production machine learning," in *The 2017 ACM International Conference on Management of Data (SIGMOD 2017)*, May 2017, pp. 1723–1726.

[14] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. M. Leung, "A survey on security threats and defensive techniques of machine learning: A data driven view," *IEEE Access*, vol. 6, pp. 12 103–12 117, February 2018.

[15] S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Cortés, "A survey on metamorphic testing," *IEEE Transactions on Software Engineering*, vol. 42, no. 9, pp. 805–824, September 2016.

[16] E. J. Weyuker, "On testing non-testable programs," *The Computer Journal*, vol. 25, no. 4, pp. 465–470, November 1982.