Requirements Engineering for Artificial Intelligence: What is a Requirements Specification for an Artificial Intelligence?

Daniel M. Berry[0000−0002−6817−9081]

Cheriton School of Computer Science, University of Waterloo
Waterloo, ON, N2L 3G1 Canada
dberry@uwaterloo.ca
https://cs.uwaterloo.ca/~dberry/

Abstract. **Context**: This article concerns requirements for an artificial intelligence (AI) that does a non-algorithmic task that requires real intelligence. **Problem**: The literature and practice of AI development does not clarify what is a requirements specification (RS) of an AI that allows determining whether an implementation of the AI is correct. **Principal ideas**: This article shows how (1) measures used to evaluate an AI, (2) criteria for acceptable values of these measures, and (3) information about the AI's context that inform the criteria and tradeoffs in these measures, collectively constitute an RS of the AI. **Contribution**: This article shows two related examples of how such an RS can be used and lists some open questions that will be the subject of future work.

Keywords: Recall and precision · Empirical acceptability criteria · Tradeoff

1 Introduction: Background and Some Related Work

The desire is to develop an artificial intelligence (AI)[1] that does a non-algorithmic *task* that requires real intelligence (RI), i.e., from a human, e.g., to recognize a stop sign in an image. In general, a task is to find *correct answers* in a space of *answers*, some of which are *correct* and the rest of which are *incorrect*. This AI might be

- a *classical* AI, which is an algorithmic attempt to simulate a human's thinking as E[2] does the task, perhaps with the help of logic, or
- a *learned machine* (LM)[3], which is the result of an instance of machine learning (ML) or deep learning, whether the LM is taught, self-taught, or both with relevant real-world (RW) data.

[1] Glossary of Non-Standard Acronyms:

HAP humanly achievable precision	RI real intelligence
HAR humanly achievable recall	RW real world
LM learned machine	ZJVF Zave–Jackson Validation Formula

[2] "E", "em", and "er" are gender non-specific third-person singular pronouns in subjective, objective, and possessive forms, respectively.

[3] a.k.a."ML component (MLC)" [18]

This article uses the term “an AI” to mean any of these possibilities, as well as any other that may be discovered or invented in the future.

It has been my observation that no AI worker expects to be able to describe an AI’s behavior completely, and everyone works around this limitation to describe an AI’s behavior in imprecise terms, such as “usually”, “probably”, “approximately”, etc., giving only empirically determined probabilities. An AI is evaluated with *vague*[4] measures, such as recall and precision. While there might be a simple specification of a task, e.g., “Return only images that contain stop signs.”, there is no actionable specification that identifies all and only images containing stop signs. Thus, there is no possibility of a formal mathematical specification. And yet, it is desired to be able to say with some certainty whether an implementation of an AI for a task does indeed do the task, at least well enough [1, 2, 9, 10, 12, 14, 17, 18, 20, 22, 24, 25].

Some have asked a key question that seems not to be satisfactorily answered in the literature [1, 13, 25].

> How does one write a requirements specification (RS), $\mathcal{S}$, for an AI, $\mathcal{A}$, for a task, $\mathcal{T}$, in a way that $\mathcal{S}$ can be used to decide whether $\mathcal{A}$ correctly implements $\mathcal{T}$, by asking whether $\mathcal{A}$ satisfies $\mathcal{S}$?

If $\mathcal{A}$ is an LM, which is a data-centric system, $\mathcal{S}$ includes the RW data with which $\mathcal{A}$ learned to do what it does [1, 2, 5, 10, 12].

2 Basic Approach

Fundamentally, an AI for a task must *mimic* humans who are using their RI to perform the task [10, 22, acknowledged Alessio Ferrari]. Lacking any complete specification of the task, we accept that what humans do in practice, while trying to avoid bias [15, 26], is correct. The mimicry will rarely, if ever, be perfect. Thus, an RS for an AI doing the task must describe this mimicry in a way that allows *measuring how well* the AI mimics humans [25]. These measures are vague and whether their values are satisfactory will not have binary, “yes” or “no”, answers. Thus, the decision about how well the AI mimics humans will be a matter of judgment. One such set of measures is *recall and precision*, measures of the frequency of correctness w.r.t. a human-determined gold set. There are other sets of measures that achieve the same objective [16]. See Section 8 about future work concerning other measures.

2.1 Zave–Jackson Validation Formula (ZJVF)

The measures are vague and not binary, and human performance in the RW is part of the decision. Thus, the truth of the claims that the evaluation criteria are met and, thus, an

[4] I.e., there is little certainty on what values of the vague measure are good and are bad. Even when there is certainty that some value is good and another value is bad, there is no certainty about what value in between is the boundary between the good and the bad.

RS is satisfied, is not logical, but is empirical, just as with the Zave–Jackson Validation Formula (ZJVF)[5],

$$\mathcal{D}, \mathcal{S} \vdash \mathcal{R},$$

which is about any computer-based system (CBS) that interacts with the RW [27, 8].

The ZJVF assumes that the RW in which a CBS operates has been divided into an environment, *Env*, and a system, *Sys*, that intersect at their interface, *Intf*, as is shown in Figure 1.

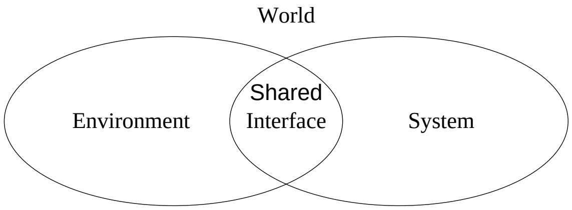

Fig. 1. The ZJVF Worldview of a CBS

The *Env* is the part of the world that is affecting and is affected by *Sys*, and *Sys* is the CBS that is desired by the stakeholders who provide the requirements. The elements of the ZJVF, $\mathcal{D}, \mathcal{S} \vdash \mathcal{R}$, where "$\vdash$" is "entailment", are assertions $\mathcal{D}$, $\mathcal{S}$, and $\mathcal{R}$:

- $\mathcal{D}$, domain assumptions, written in the vocabulary of *Env*, is what *Sys* is allowed to assume about *Env* in its execution in *Env*;
- $\mathcal{S}$, specification, written in the vocabulary shared by *Env* and *Sys*, i.e., the vocabulary of *Intf*, is a description of the behavior of *Sys*; and
- $\mathcal{R}$, requirements, written in the vocabulary of *Env*, is a description of the stakeholders' requirements for *Sys* in terms of *Sys*'s effect on *Env*.

The validation formula, $\mathcal{D}, \mathcal{S} \vdash \mathcal{R}$ says that *Sys* meets its requirements in *Env* if the empirical truth of $\mathcal{D}$ in *Env* in the RW is enough for $\mathcal{S}$ to entail $\mathcal{R}$.

In this diagram, the RW for an AI is *Env*, and the learning data, *LD* with which an LM learns is in *Intf*. $\mathcal{D}$ must include that *LD* is true and *representative* of the RW. Thus, even though

1. the AI's code, which is written in a programming language, is a formal object, and the truth of the claim that the code implements $\mathcal{S}$ is logical, and
2. the whole formula looks formal,

since $\mathcal{D}$ and $\mathcal{R}$ are about the RW, the truth of the whole formula is empirical. In addition, if $\mathcal{S}$ is about an LM then $\mathcal{S}$ includes *LD*, data about the RW, then the truth of $\mathcal{S}$ becomes empirical as well.

[5] The $\mathcal{S}$ in the ZJVF could very well be the $\mathcal{S}$ mentioned in the key question that seems not to be satisfactorily answered in the literature, at the end of Section 1. So it's OK that they have the same typeface!

2.2 Running Examples

As running examples, this article uses two different AIs, $A1$ and $A2$, for the task of finding stop signs in images, in two different contexts that impose different needs on the measures. Each AI is to classify each input image as to whether or not the image has at least one stop sign, and is to output only those images that do. The difference between $A1$ and $A2$ is in the way the outputs are used. $A1$ finds the images that contain stop signs in order to produce a set of images, with which to train $A2$ to identify stop signs in real time for an autonomous vehicle (AV). This article describes these two different AIs with the same base functionality to demonstrate the necessity of including in an RS for the AI, the context of the AI's use. The use of the same algorithm and the same RW training data for these AIs would yield the same recall and precision values, not distinguishing the AIs. Only the context distinguishes them and allows determining whether the recall and precision values are acceptable for the AI's use.

2.3 Plan for the Rest of the Article

This article tries to show that *any* set of measures that is used to evaluate an AI in an attempt to convince the AI's stakeholders that the AI is what they want [23] can be the basis of an RS of the AI *if* added to this basis is all the information from the AI's context that the stakeholders need about the meanings of the values of the measures, to be able to decide whether the AI is satisfactory for their needs.

In the rest of this paper, Section 3 reminds the reader about recall, precision, and summarization. Section 4 describes the two AIs, $A1$ and $A2$, and how they may be evaluated, allowing Section 5 to abstract to a general framework for an RS for an AI. Section 7 summarizes related work, and Section 8 points to future work.

3 Recall, Precision, Summarization

In the interest of conserving space in this article, this article merely reminds the reader of the meanings of *recall*, *precision*, and *summarization*, which are described fully elsewhere [3][6]. For an AI, A

- recall (R): percentage of the correct answers that are returned by A,
- precision (P): percentage of the answers returned by A that are correct, and
- summarization [7] (S): percentage of the input to A that are removed in the output that A returns, i.e., $(100\% - (\frac{\mathit{size(output)}}{\mathit{size(input)}}))$.

Informally, the output of an AI is correct if it has *all* and *only* correct answers. R and P are the two sides of "*all* and *only*": R measures how close to *all* correct answers are in the output. P measures how close to *only* correct answers are in the output. S measures

[6] It was a total surprise that the cited work was so applicable to RSs for AIs.

[7] ALERT: This summarization is not what is usually called "summarization" in the context of AI. It is not what the AI does, but a measure about the sizes of the input and output to the AI. So please look carefully at what the definition says.

how much of the task that the AI is supposed to do is done and is not left to be done by humans.

To clarify the measures, the importance of context, and the importance of summarization, consider an application of one of the running examples, $A1$, to a set of 1000 images, of which 200 contain stop signs. With this distribution of correct answers, images with a stop sign, among the answers, the images, in a manual search for correct answers, on average 5 answers will have to be examined to find one correct answer. Thus, finding a correct answer costs 5 times what examining an answer costs. Suppose that $A1$ returns 400 images of which 190 truly have stop signs. Then,

$R = \frac{190}{200} = 95\%$,
$P = \frac{190}{400} = 47.5\%$, and
$S = 100\% - \frac{400}{1000} = 60\%$.

These particular measure values are not bad, particularly if the average human has poorer than 95% recall in the same task. Because the output of $A1$ is being used to train $A2$, it is essential to get as close as possible to having *all* and *only* images that contain stop signs. Because $P = 47.5\%$ means that more than half of $A1$'s output is false positives, $A1$'s output must be manually searched, i.e., *vetted*, to find them and remove them. The 60% summarization says that the manual vetting search of the only 400 images returned by $A1$ will be considerably faster than a manual search of the original 1000 images. Thus, the poor precision of 47.5% does not matter that much, because the tedium of a manual search has been cut by 60%. As observed by a reviewer of the conference paper that is based on this article, any way of ensuring that vetting is fast is OK, e.g., that a human's correctness decision for an item in the AI's output is considerably faster than for an item in the AI's input [3].

4 Evaluation of $A1$ and $A2$ with the Measures

If we decide to use recall and precision as the basis for the evaluation and, thus, specification of an AI, then the process of determining if an implementation meets the specification involves (1) evaluating and comparing the recall and precision of the AI and of humans doing the same task and (2) using the context of the task, which is thus part of the specification, as the basis for deciding what the comparison means.

For $A1$ and $A2$, each AI is evaluated by its R and P, with respect to a manually developed gold set of classified images. Each human expert in the domain of the AI that participates in developing the gold set computes er own R and P, and the averages of their R and P values are

- the *humanly achievable recall* (HAR) and
- the *humanly achievable precision* (HAP)

of the stop-sign recognition task. Each of these HAR and HAP is probably about 99%[8].

[8] This claim needs to be tested empirically, but probably there are very accurate data at www.captcha.net.

One possibility is to require an AI for a task to at least mimic people doing the same task. Otherwise, especially for a life-critical task, we're better off leaving the task to humans [4]. So, one possibility for an AI for a task is

- for the AI's R to achieve or beat the task's HAR and
- for the AI's P to achieve or beat the task's HAP.

In the case of $A2$, achieving or beating HAR and HAP is acceptable; accidents are inevitable, particularly if humans are doing the task. If $A2$'s R and P achieve or beat the task's HAR and HAP, then $A2$ will have no more accidents than does a human doing the task. While no accident is good, society can accept an AI's doing this task in this circumstance.

In any AI for a life-critical task, regardless of how important a high P is, a high R is very critical. Finding *all* correct answers is often *very* necessary. Lives depend on doing so. However, if achieving high R is important, there may be *no* choice but to accept low P, because in many algorithms for tasks that require RI, recall and precision trade off. That is, a higher R can be achieved only at the cost of lowering P, and vice versa.

For each of $A1$ and $A2$, achieving or beating the task's HAR is essential. However, for $A1$, a low P means that there are lots of false positives among the output of $A1$. Fortunately, for $A1$'s specific context, these false positives are not really dangerous, because there is plenty of time for vetting to find the false positives and remove them from the output. However, lots of false positives among the output of $A1$ can discourage the human vetters. If S is high, then the vetters can be reminded that manually vetting $A1$'s output is a lot faster than manually searching $A1$'s entire input. Unless S is actually zero, $A1$ *does* reduce the manual searching that needs to be done. In a vetting context, the R and P of the AI is determined only *after* the vetting, because vetting does generally improve P. In the end, for $A1$, if the R after vetting beats the task's HAR, and the time to vet $A1$'s output is less than the time to do the classification task manually, then $A1$ is considered to meet its requirements. After all, since the task of $A1$ is essential, the alternative to running $A1$ is to do the task completely manually at the cost of a lot *more* tedious, *boring* grunt work!

$A2$ runs in an AV, making vetting impossible, because there is not enough time between recognition of a stop sign and the need to press the AV's brakes. Also, the AV would not be autonomous if a human vetter were present in the vehicle. Therefore, low P means lots of unnecessary stops by the AV, that could very well lead to dangerous rear-end collisions, caused by the surprised drivers of the vehicles following the AV! Therefore, for $A2$, low P is definitely not tolerable, and reusing $A1$ as $A2$ is not acceptable. Another $A2$ must be found that makes both R and P high enough to achieve or beat the task's HAR and HAP [6].

This example has suggested one particular set of measures, — R, P, and S — and one particular set of criteria — R's and P's achieving HAR and HAP, possibly with the help of vetting assisted by a high S. However, *any* set of measures and *any* criteria that make sense to an AI's stakeholders can be used as the RS for the AI.

Even in a vetting situation, there are different contexts. For example, if an AI is to identify the image of a cancerous tumor in a radiograph for one patient, then the time to vet is not a factor at all. The examining domain expert can take all the time E needs to make correct decision. If, however, the AI is to identify the same for a large number of patients, then the total time to vet is an important factor in deciding whether the AI is satisfactory. The total time to vet depends on the time to vet one item and on the number of items to vet, and the number of items to vet depends on S, the summarization, of the AI.

My son's start up is developing an AI that will make life-critical medical decisions from data, decisions that are difficult for humans to make because of the large volume of data that are relevant. We want both R and P to be 100%, or at least beating the task's HAR and HAP, for the patients' sakes. The start up has several high-P AIs, each with a P very close to the task's HAP, but none has an R better than 50%. Fortunately, the region of the recall of each pair of AIs overlaps only a little. So, the recommendation is to try running them all to see if the union of their outputs has an R that beats the task's HAR! After all, computers and running software are cheap.

5 What an RS for an AI is

It is now clear that an RS for an AI needs more than just whatever measures $M_1, \dots,$ and M_n are used in evaluating the AI. The RS needs also criteria for acceptable values of these measures, e.g.,

- minimum, or maximum, threshold values of $M_1, \dots,$ and M_n, which may be the humanly achievable values of $M_1, \dots,$ and M_n for the AI's task, with which to compare the AI's $M_1, \dots,$ and M_n values, respectively;
- the relative importance of the individual measures $M_1, \dots,$ and M_n to help evaluate any needed tradeoff between $M_1, \dots,$ and M_n [6];
- in a case in which vetting is possible or required, (1) the S of the AI and (2) the times for a human to decide the correctness of an item in the AI's input and in the AI's output; and
- any data, e.g., training data, that are needed for the AI to function correctly.

Calculating the relative importance of, and thus the tradeoffs between, the measures $M_1, \dots,$ and M_n in the context of the AI requires a *full* understanding of the context in which the AI is being used, including the cost of achieving a high value in each of the individual measures $M_1, \dots,$ and M_n, in the context [6]. Non-functional requirements will help define the context and decide the tradeoffs [9, 25].

Finally, the decision of whether the AI satisfies its RS and meets its requirements will involve engineering judgement and evaluation of tradeoffs in the AI's context, and will *not* be a simple "yes" versus "no" decision, because of all of the vague elements in the RS. The RS for an AI is as vague as are fitness criteria for vague qualitative, non-functional requirements, e.g., "fast response time" or "friendly user interface". For examples:

1. What should be done if the value of any measure *just misses* its threshold while all the others beat their thresholds?
2. How *critical* must the task be in order that an acceptable alternative to an AI that does not satisfy its RS is doing the task manually?
3. How fast must vetting be for vetters to *tolerate* having to vet?

Questions like these can interact in an engineering way. For example, what should be done in the situation in which the task is *only fairly critical*, the AI *just misses achieving* the task's thresholds, and vetting is *somehat slow*?

To place this form of RS in the milieu of the ZJVF, consider each of $A1$ and $A2$. For each,

- the non-actionable specification of the shared task of $A1$ and $A2$

  Return only images that contain stop signs, as well as a human would.

  would be the $\mathcal{R}$ in $\mathcal{D}, \mathcal{S} \vdash \mathcal{R}$;
- all the measures and criteria described in Section 4 as being part of the RS for $A1$ or $A2$ would be the $\mathcal{S}$ in $\mathcal{D}, \mathcal{S} \vdash \mathcal{R}$; and
- assertions about all facts about the RW that are needed for the entailment of $\mathcal{R}$ by $\mathcal{S}$ would be the $\mathcal{D}$ in $\mathcal{D}, \mathcal{S} \vdash \mathcal{R}$;

Observe how the criteria embedded in $\mathcal{S}$ are directed at ensuring that the AI do its task as well as a human would.

6 RE for AI

All of this information is what *requirements engineering (RE) for an AI* must elicit or invent, and therefore, potentialy all of RE's methods must be applied to elicit or invent the context and its description for inclusion into the RS. All of the judgements and tradeoffs mentioned in the last paragraphs of Section 5 are therefore parts of the RE for AI [11]. In this sense, RE for AI is not very different from RE for any complex CBS that interacts with the RW.

7 Related Work

Most of the related work is cited at any point in this article where an observation or contribution made by the work is mentioned.

Salay and Czarnecki observe that the ISO 26262 standard does not prescribe a single complete specification for partially or fully autonomous vehicles, describing only a collection of specifications for development processes and individual properties of an AV [20]. They address the difficulties, including some mentioned in Sections 1 and 2 of this article, with these specifications by providing improvements to each of the specifications of the standard. The RS framework suggested by this article will need to incorporate their improvements. See Section 8.

Kästner observes that the engineering of the RW training data to yield the desired behavior corresponds to RE rather than implementation of the resulting LM [12]. Checking with the customer that the training data yields the correct behavior is validation. Thus, these training data end up being part of the specification of the LM.

There are methods to test whether an AI does what it is supposed to do [2, 21, 28]. Implicitly, whatever the test data test are the requirements of the AI.

Ribeiro, Ribeiro, and Castro conducted a systematic literature review of the topic of RE for AVs [19].

Rahimi *et al* use “machine learned component (MLC)” to describe what this article terms “LM” [18]. They too attempt to explain what a requirements specification for a MLC is or at least to improve the process of specifying a CBS that contains an MLC. Their approach “extracts a universally accepted benchmark for hard-to-specify concepts (e.g., ‘pedestrian’) and can be used to identify gaps in the associated dataset and the constructed machine-learned model.”

They observe that the typical specification of many a MLC is not what this article calls “actionable”:

> For example, the requirement for the automated pedestrian collision avoidance system might specify that “the position of the pedestrian should be detected within an accuracy of 0.5m”. However, decomposing such high level specifications to lower-level verifiable ones is difficult, if not impossible.

while citing [22].

Their approach involves use of component-level specification to define the behavior of an MLC. They facilitate unambiguous specification of MLCs by building a benchmark on the Web for domain concepts that are hard to specify. They offer S = “The pedestrian detector component shall be able to detect pedestrians on foot, on a scooter and on a wheelchair” as a suitable specificatioon for a pedestrian detector component. They then claim that are able to systematically verify an MLC against the set of derived specifications: “For example, we can verify whether the component is able to correctly classify a pedestrian on a wheelchair or a scooter.”

It is not clear to me either

- how they can derive S from “ a pedestrian detector component” or
- how S is more actionable, decomposable, and verifiable than “the position of the pedestrian should be detected within an accuracy of 0.5m”.

Perhaps, in getting the paper down to 4 pages, they have left out essential information. Nevertheless, I do not see anything in the paper that can serve as an actionable specification for a LM.

There is a lot of somewhat related work in the proceedings of the AIRE Workshops (https://ieeexplore.ieee.org/xpl/conhome/1803944/all-proceedings) and of the RE4AI Workshops (http://ceur-ws.org/Vol-2584/, http://ceur-ws.org/Vol-2857/). Papers from these workshops that address the topic of this article are cited in this article.

8 Future work

Section 5 shows only a first attempt at abstracting from what was learned from the running example to a general framework for RSs for AIs. The details of this framework changed a lot prior to submission of this article and as a result of the reviewers’ comments. It is, thus, clear that the main future research will be to examine more AIs to understand their measures, criteria, and contexts in the hopes of arriving at a statement of the framework that works for all AIs. Also, Section 4 considered only *one* possible meta-level requirement, that an AI for a task at least mimic people doing the same task.

Are there other possibilities? These need to be explored. Of course the meta-level requirement, which is used to inform the criteria for the measures, becomes a requirement for the AI. Nevertheless, the basic idea remains: The RS for an AI consists of a description of all measures and criteria plus all information about the AI's context of use that are necessary for the AI's stakeholders to decide if the AI meets their needs.

Some specific topics include:

- Are there measures, other than recall and precision, on which an RS for an AI can be based? Examples include
  1. other measures calculable from a confusion matrix [16],
  2. especially, sensitivity (the same as recall) and specificity (recall of the true negatives), which are used in medical situations in which a true negative is just as important as a true positive [7], and
  3. interrater agreement between the AI and some humans using their RI.
- What is the role in an RS of the representativeness of the data with which an LM is trained in the RS of the LM [1, 2, 5, 12]?
- What is the role in an RS of existing industrial standards such as the ISO 26262 standard for AVs [20]?

Acknowledgments

I benefited from voice and text discussions with Tiago Amorim, Krzysztof Czarnecki, Nancy Day, John DiMatteo, Patrick Ebel, Alessio Ferrari, Jannik Fischbach, Julian Frattini, Vijay Ganesh, Andrea Herrmann, Hans-Martin Heyn, Jeff Joyce, Mersedeh Sadeghi, Aaron Schlütter, Davor Svetinovic, John Thistle, Richard Trefler, and Andreas Vogelsang. I thank the anonymous reviewers of the conference paper derived from this article for their suggestions.

Daniel Berry's work was supported in part by a Canadian NSERC grant NSERC-RGPIN227055-15.

References

1. Ahmad, K., *et al*: What's up with requirements engineering for artificial intelligence systems? In: IEEE 29th RE. pp. 1–12 (2021)
2. Ashmore, R., *et al*: Assuring the machine learning lifecycle: Desiderata, methods, and challenges. ACM Comp. Surv. **54**(5), 111 (2021)
3. Berry, D.M.: Empirical evaluation of tools for hairy requirements engineering tasks. EMSE **26**(6), 111 (2021)
4. Berry, D.M., *et al*: The case for dumb requirements engineering tools. In: REFSQ. pp. 211–217 (2012)
5. Chuprina, T., *et al*: Towards artefact-based requirements engineering for data-centric systems. In: REFSQ-JP 2021: RE4AI (2021)
6. DiMatteo, J., *et al*: Requirements for monitoring inattention of the responsible human in an autonomous vehicle: The recall and precision tradeoff. In: REFSQ-JP 2020: RE4AI (2020)
7. Greenfield, Y.: Precision, recall, sensitivity and specificity (2012), https://uberpython.wordpress.com/2012/01/01/precision-recall-sensitivity-and-specificity/
8. Hadar, I., *et al*: The inconsistency between theory and practice in managing inconsistency in requirements engineering. EMSE **24**(6), 3972–4005 (2019)

9. Horkoff, J.: Non-functional requirements for machine learning: Challenges and new directions. In: IEEE 27th RE. pp. 386–391 (2019)
10. Hu, B.C., *et al*: Towards requirements specification for machine-learned perception based on human performance. In: IEEE 7th AIRE. pp. 48–51 (2020)
11. Ishikawa, F., Matsuno, Y.: Evidence-driven requirements engineering for uncertainty of machine learning-based systems. In: IEEE 28th RE. pp. 346–351 (2020)
12. Kästner, C.: Machine learning is requirements engineering (2020), https://medium.com/analytics-vidhya/machine-learning-is-requirements-engineering-8957aee55ef4
13. Kostova, B., *et al*: On the interplay between requirements, engineering, and artificial intelligence. In: REFSQ-JP 2020: RE4AI (2020)
14. Kress-Gazit, H., *et al*: Formalizing and guaranteeing human-robot interaction. CACM **64**(9), 78–84 (2021)
15. Mehrabi, N., *et al*: A survey on bias and fairness in machine learning. ACM Comp. Surv. **54**(6) (2021)
16. Mishra, A.: Metrics to evaluate your machine learning algorithm (2018), https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
17. Parnas, D.L.: The real risks of artificial intelligence. CACM **60**(10), 27–31 (2017)
18. Rahimi, M., *et al*: Toward requirements specification for machine-learned components. In: IEEE 27th RE Workshops (REW). pp. 241–244 (2019)
19. Ribeiro, Q.A.D.S., *et al*: Requirements engineering for autonomous vehicles: A systematic literature review. In: 37th ACM Symp. on Applied Computing (SAC '22) (2022)
20. Salay, R., Czarnecki, K.: Using machine learning safely in automotive software: An assessment and adaption of software process requirements in ISO 26262 (2018), https://arxiv.org/abs/1808.01614
21. Schmelzer, R.: How do you test AI systems? (2020), https://www.forbes.com/sites/cognitiveworld/2020/01/03/how-do-you-test-ai-systems/
22. Seshia, S.A., *et al*: Towards verified artificial intelligence (2020), https://arxiv.org/abs/1606.08514v4
23. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, New York, NY, USA (2014)
24. Valiant, L.: Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World. Basic Books, New York, NY, USA (2013)
25. Vogelsang, A., Borg, M.: Requirements engineering for machine learning: Perspectives from data scientists. In: IEEE 27th RE Workshops (REW). pp. 245–251 (2019)
26. Wing, J.M.: Trustworthy AI. Communications of the ACM **64**(10), 64–71 (2021). https://doi.org/10.1145/3448248
27. Zave, P., Jackson, M.: Four dark corners of requirements engineering. TOSEM **6**(1), 1–30 (1997)
28. Zhang, J., Li, J.: Testing and verification of neural-network-based safety-critical control software: A systematic literature review. IST **123**, 106296 (2020)