

## Decision Letter (CACM-22-06-4286)

**From:** eic@cacm.acm.org

**To:** dhaminda.abeywickrama@bristol.ac.uk

**CC:** dhaminda.abeywickrama@bristol.ac.uk, amel.bennaceur@open.ac.uk, greg.chance@bristol.ac.uk, y.demiris@imperial.ac.uk, a.kordoni@lancaster.ac.uk, mark.levine@lancaster.ac.uk, l.moffat1@lancaster.ac.uk, luc.moreau@kcl.ac.uk, mohammad.mousavi@kcl.ac.uk, B.Nuseibeh@open.ac.uk, s.ramamoorthy@ed.ac.uk, jan.ringert@uni-weimar.de, j.wilson@bristol.ac.uk, shane.windsor@bristol.ac.uk, kerstin.eder@bristol.ac.uk

**Subject:** Communications of the ACM - Decision on Manuscript ID CACM-22-06-4286

**Body:** 05-Nov-2022

Dear Dr. Abeywickrama:

Manuscript ID CACM-22-06-4286 entitled "On Specifying for Trustworthiness" which you submitted to the Communications of the ACM has been reviewed. The reviewer(s)' comments are included at the bottom of this email.

The reviewer(s) have recommended publication but suggested substantial revisions to your manuscript. Therefore, I invite you to respond to the reviewer(s)' comments and revise your manuscript.

To revise your manuscript, log into <https://mc.manuscriptcentral.com/cacm> and enter the Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Then, under "Actions," click on "Create a Revision." Your manuscript number will be appended, denoting a revision.

You cannot change the originally submitted version of the manuscript. Instead, revise a saved copy of the manuscript and resubmit the new version. Please highlight the changes to your manuscript using the track changes mode in MS Word or by using bold or colored text. <b>Your revision must continue to adhere to the CACM Author Guidelines.</b>

You should upload the new version and submit it through the Author Center when the manuscript is revised.

When submitting your revised manuscript, you will be able to respond to the comments made by the reviewer(s) in the space provided. In addition, you can use this space to document other changes to the original manuscript. To expedite the processing of the revised manuscript, please be specific in describing changes made in response to the reviewer(s)' comments.

Because we are trying to facilitate the timely publication of manuscripts submitted to Communications of the ACM, you should upload your revised manuscript as soon as possible.

If you cannot submit your revision in a reasonable amount of time, we may have to consider your paper as a new submission. Please see deadlines on your ManuscriptCentral Author Dashboard. If you need a slight extension, please contact the Editor-in-Chief.

Once again, thank you for submitting your manuscript to the Communications of the ACM and I look forward to receiving your revision.

Sincerely,  
James R. Larus  
Editor-in-Chief, Communications of the ACM (CACM)  
Professor, EPFL, Lausanne Switzerland

Co-Chair: Co-Chair, Contributed  
Comments to the Author:  
(There are no comments.)

Associate Editor: Cleland-Huang, Jane

Comments to the Author:

The reviewers have mixed opinions about this paper -- one finds it exciting and highly topical and the other, not so much. Both reviewers are quite confused about the purpose and focus of the paper, and the problem that it is trying to address. They are not convinced by the paper's claims.

They both ask for major improvements. In particular, should you choose to submit a major revision please do the following:

1. Clarify the major contributions of the paper
2. Explain why the specific domains/systems were selected
3. Explain why the trust issues in the domain of AS are different (if they are) from those of other safety-critical domains.
4. Make sure that the claims about the intellectual challenges are well justified.

As the reviewers have both recommended major revision, I'm making this recommendation; however, the paper needs a solid rethink and needs to clearly articulate the rationales behind the claims and categorizations in the paper.

Reviewer: 1

Recommendation: Major Revision

Comments:

This paper is concerned with the concept of trust and trustworthiness of Autonomous systems (AS). The authors present some of the challenges for specifications in several autonomous systems in order to explore how should specifications for these systems be developed. The article is exploratory in nature and reads more like a philosophical narrative, a scholarly argument, a position statement, or a partial roadmap for research directions rather than a rigorous research article with practical applications.

Trustworthiness of AS in the introduction is defined as when the design, engineering, and operation of these systems generate positive outcomes and mitigate outcomes that can be harmful. This definition seems to be misaligned with another statement in the introduction about trust: "human may trust an AS to perform its actions, if it demonstrably acts in an effective and safe manner". It is not clear what "effective" and "safe manner" means in this context and how they can be measured.

While I found figures 1 and 2 a reasonable encapsulation of the core content of the paper, I am not really sure what the major contributions of this paper are and am not convinced if this paper could attract much interest or trigger any debate within the Autonomous Systems research or practice community.

Many of the specification challenges described are similar to the challenges faced in non-autonomous systems such as high assurance or safety-critical systems. I would have liked to see the authors compare and contrast these specification challenges between AS and non-AS more rigorously to strengthen the argument about the new features that warrant a different approach. Also, many of the intellectual challenges for the autonomous systems community are very similar and some almost identical to the challenges identified and addressed by the multi-agent systems in particular and the Artificial Intelligence (machine learning) community in general. For example, the notions of trust and trustworthiness have been well researched and written about under AI Ethics principles and Responsible AI research work. E.g. Zhu, L., Xu, X., Lu, Q., Governatori, G., Whittle, J. (2022). AI and Ethics—Operationalizing Responsible AI. In: Chen, F., Zhou, J. (eds) Humanity Driven AI. Springer

I am not sure why "AI in healthcare" has been singled out from numerous applications of AI.

Question about explainability that authors ask (what is an actual explanation and how should an explanation be constructed?, What is the purpose of an explanation? What is the audience of an explanation? What is the information that it should contain?) have all been asked in numerous publications and many answers and solutions proposed in specific contexts for AI (which is arguably the most popular and the most used autonomous system).

Specifications are traditionally known to be the artefact that is the outcome of requirements engineering (RE) in any systems development. I find it very surprising that in the vast majority of the frameworks, lifecycles, methods, and techniques proposed and used for the development of AS (AI in particular) today, RE is largely absent and references to "requirements" are very scarce. Perhaps, the need for RE is not recognised and writing specifications as we know it is not considered to be a particularly useful way of spending valuable development time anymore.

As a side, I find the acronym (ASS) kind of strange for a scientific article, may I suggest using just (AS) instead?! ;-)

Additional Questions:

Reviewer: 2

Recommendation: Major Revision

#### Comments:

The paper addresses the interesting problem of analyzing what is needed to include trustworthiness requirements while specifying autonomous systems (ASs). This is a very interesting and timely topic. Highlighting the challenges the ASs community has to face in specifying trustworthiness as a requirement can support the community in developing organized and effective solutions.

However, while some of the identified challenges are clearly relevant and to be faced to specify trustworthiness, in its current status, the paper is very confusing and (1) does not clearly explain the details behind the different classifications (e.g., chosen domains) and selections (e.g., priorities), (2) does not motivate while (only) certain specification challenges were selected for specific domains (and not for others), and (3) does not clarify how/from where the intellectual challenges were identified. In addition, while some of the challenges are related to trustworthy ASs, others are more general challenges (faced independently from the need of specifying trustworthiness or adaptation).

In general, while the paper is an enjoyable reading, the presentation does not suggest that enough rigor was used to identify the domains and the questions. Thus, I think, it does not convince the reader that the identified challenges are (all) the relevant ones to guide the ASs community. Specific comments/questions follow.

#### **\*\*Domains and challenges associated/selected for each domain\*\***

The authors identified 6 domains and initially reasoned on the main challenges associated with each of them.

I am not sure if I missed some relevant references or explanations, but it is not clear to me the origin of this classification. How have the domains been identified? It seems to me that some of the categories identify the domains for/in which ASs are used (e.g., emergencies, healthcare), and others group ASs depending on the kind of devices they are (e.g., drones). This blurry categorization creates a lot of overlap among the categories and this makes it unclear whether certain aspects are relevant for one category and not the other. Here are some specific questions/comments:

- Pag 2, col 1, 56- col 2, 5: Isn't this a general problem of ASs in open environments where the other actors are not necessarily autonomous systems themselves (rather than of this specific domain)?
- Pag 2, col 2, In 19-22: I fail to understand how this differs from the automated driving systems. These all work in open environments. Also, I fail to see how to group in a category all the autonomous systems in the domain of emergency/disaster recoveries, as they all deeply vary one from the other.
- Pag 2, col 2, In 27-30: similar concern.
- Pag 2, col 2, In 44-50: while the human mental state is relevant only for this domain?
- AI in healthcare represents a type of ASs (AI) and a domain -healthcare (domain that is also considered in the previous group - human-robot interaction).
- pag 3 col 2 In 33-38, If you consider multiple autonomous vehicles in an environment... how is this different?
- Uncrewed aerial vehicles can be used in different domains (some mentioned as other domains).

#### **\*\*Intellectual Challenges\*\***

It is unclear to me how the intellectual challenges have been derived/selected and how they relate to the challenges specific to each domain. I want to clarify that I do agree that some of these challenges need to be addressed by the ASs community that needs to provide sustainable solutions to them. However, there is no rationale, pattern, or explanation to show the readers why these are the challenges that have been identified. In addition, the challenges are at very different "levels". Some are very specific to trustworthiness for ASs, some to ASs, and some are general to software development. Finally, the identified research directions are not motivated. Here are some specific questions/comments:

The way in which the problem of evolution is presented in pag 6-7 (evolve specifications) does not seem to be specific to ASs or trustworthy ASs.


- The problem of incompleteness has been faced also with the use of partial models (see Chechik's, Gurfinkel's, and Uchitel's work) which can be impactful also in this context.
- Most of the considerations done about competing demands (pag 8 col 1) seem to be independent from trustworthiness (with the only exception of paragraph 38-46).

#### **\*\*Other comments\*\***

- The abstract does not clearly explain the need for this work. Specifically, it is not clear what is the actual problem with existing specifications and how the contribution correlates with this problem. The abstract should set the tone for the rest of the paper to propose the problem and clearly state that the highlighted challenges are proposed to support the community in overcoming the limitations of the current solutions.
- Pag 1, col 2, In 54: Is "inspiring" the goal? The specifications are used by technical validation and verifications techniques (In 48-50), so the specifications need to contain all the relevant information to be able to analyze the concept of trustworthiness of interest. I do not understand this paragraph concerning the previous one. The user will not read the specifications but will rely on the result of the analysis. This point needs to be clarified.
- On page 1, col 47-49, you say that there are techniques to prove trustworthiness, probably you should say that they try and something is missing. Otherwise, you are saying here something in opposition (pag 2, col 1, In 9-14).
- Pag 2, Col 1, In 38-49: Are these rules uniquely interpreted by people? Are they supposed to be? The key point is this difference and its perception.
- The conclusion mentions the difference between the human ability to grasp "rules" that are difficult instead for ASs (as, I believe, they intend the case of street rule and autonomous driving). This aspect is very important, but not really developed in the paper. Especially, when talking about rules on purpose developed to leave space for interpretation.

Additional Questions:

**Date Sent:** 05-Nov-2022

 Close Window