==========

S&P 2020 Review #732A

Paper #732: Dangerous Skills Got Certified: Measuring the Trustworthiness of Amazon Alexa Platform

---

Overall merit: 4. Weak accept - While flawed, the paper has merit and we should consider accepting it.

===== Brief paper summary (2-3 sentences) =====

This paper measures the trustworthiness of the Amazon Alexa platform. In particular, it shows that skills violating Amazon content guidelines can be easily certified, that the outcome of the certification process is inconsistent, that there are several certified skills that violate Amazon
content guidelines, and that skills with names similar to existing skills can be mistakenly enabled (by voice) if the number of reviews of such skills is manipulated.

The paper also shows initial evidence that the Google Assistant platform is partially affected by the same problems.

Finally, the paper proposes some strategies that voice assistant providers can implement to improve the trustworthiness of the certification process.

===== Strengths =====

- The paper is interesting, original, and timely since there are current many concerns about voice assistants.

- The paper is rich in experiments and analyses that help the reader to understand the trustworthiness from different perspectives: they managed to get hundreds of so-called dangerous skills certified, they found existing dangerous skills, and they measured user perception of trustworthiness through a user study.

===== Weaknesses =====

- Some of the most important results of the paper are unsurprising since skills backends are black boxes from the voice assistant provider perspective.

- The work would feel more interesting and complete if it addressed the more general problem of trustworthiness on voice assistance platforms without being Alexa specific (the Google Assistant analysis is just preliminary).

- Some of the results and suggestions are based on speculative deductions.

===== Detailed comments for the author(s) =====

At the beginning of the paper, when giving context about how skills work, I recommend emphasizing the fact that skills backend code is a black box for the certification process and the intuitive implications of that: i.e., that skills may be vulnerable to changes in the backend (as already shown in cited related work [10,14]). Similarly, it is also intuitive that no approach (manual or automatic) can thoroughly explore the behavior of a skill after any possible sequence of invocations (due to the black box property), therefore it is not surprising that the more difficult it is to trigger the malicious

behavior of a skill, the more difficult it is for the certification process to spot such malicious behavior. The authors eventually have all their skills approved by increasing this difficulty.

The paper contains a preliminary analysis of the Google Assistant platform; however, the number of skills tested is too small to make any proper generalizations. From the results shown it looks like the Google Assistant certification process is more trustworthy than Alexa. Having a comparable set of experiments may help support these generalizations and better understand if/why the Google Assistant certification is better than Alexa.

It is not clear why, if a developer account is hosting a malicious skill, then a different developer account under the same AWS account should raise suspicion (end of page 6).

The last paragraph of section IV.C speculates that if skills are approved at night, they may be outsourced in other countries. This is possible, but cannot be verified, and also I am not sure how it is relevant: even if true, I don't see why certification teams outside the US are necessarily trained poorly compared to US ones.

It is also not clear why the lack of policy details is necessarily a bad thing. The paper states that both Amazon Alexa's and Google assistant policies prohibit "promoting violence", then Google details specific examples of "promoting violence", while Amazon Alexa's does not. This does not imply that Amazon Alexa's policy gives more freedom, but I agree that it can be more ambiguous.

Across the whole paper the authors repeatedly say that the certification process is "disorganized", "improper", and other similar adjectives. The motivation given includes the different outcomes that similar certification requests generate. As the authors also suspect, it is quite possible that the process is manual, and different apps are analyzed by different people, with some degree of unavoidable subjectivity, skill level, and luck in finding the malicious behavior. However, unless the authors propose a way to ensure that these subjective decisions are always exactly the same (hardly), the authors need to avoid such extreme adjectives.

The last paragraph of section V.A says that "the prevalence of negative reviews reflects the existence of policy-violating skills". This is speculation: there can be many reasons why a skill does not work as the user expects, thus leading to a negative review, which is not necessarily due to a policy violation. Think of bugs, performance problems, user error, etc.

The end of section V.D discusses some strategies to improve the trustworthiness of the skill certification process. However, the discussion lacks pros and cons of each strategy, especially their feasibility/cost in the current context and what system assumptions have to be changed to make them possible. I suggest addressing these concerns.

Minor comments:

Page 4: "if it's invocation" -> "if its invocation"
Page 5:
"account were used" -> "accounts were used"
"then ends the session" -> "then end the session"
"skill was used built" -> "skill was built"

Fig 3: this is too close to the text, please add vertical space under the image
Page 12: "names that has" -> "names that have"
Page 13: "if it is broken" -> "if they are broken"

==========

S&P 2020 Review #732B

Paper #732: Dangerous Skills Got Certified: Measuring the Trustworthiness of Amazon Alexa Platform

---

Overall merit: 4. Weak accept - While flawed, the paper has merit and we should consider accepting it.

===== Brief paper summary (2-3 sentences) =====

This paper presents a series of experiments to demonstrate that developers can fairly easily have skills certified by Amazon that violate various policies. They conducted a smaller study on Google and found that Google caught some of the violations that Amazon did not catch, but still certified some skills with violations. The authors also emphasize that developers can make policy-violating changes after certification that also don't seem to get caught.

===== Strengths =====

- Practical demonstration of an important problem in widely-used systems
- Not just a narrow exploit, demonstrated with a large number of cases
- Includes good discussion of ethics, implications, etc.

===== Weaknesses =====

- There is an awful lot here, including many studies, so descriptions of some studies are rather abbreviated.
- Figure 2 doesn't add much value. I would rather see a better overview of all the different studies that are being presented.

===== Detailed comments for the author(s) =====

I think this is an interesting and informative paper that is very practical and relevant. My main complaint is that the authors may be trying to do too much in one paper. Perhaps this could be addressed with a table or figure up front that presents and overview of what is going to be presented.

==========

S&P 2020 Review #732C

Paper #732: Dangerous Skills Got Certified: Measuring the Trustworthiness of Amazon Alexa Platform

---

Overall merit: 4. Weak accept - While flawed, the paper has merit and we should consider accepting it.

===== Brief paper summary (2-3 sentences) =====

The paper explores the process by which third-party skills are certified, and whether the Alexa platform performs due diligence in studying the privacy behavior of these skills. The paper finds that there are gaps in this process by successfully registering 132 "privacy-violating skills". The paper also identifies skills in the Amazon skills store that have this type of vehavior.

===== Strengths =====

Alexa Skills are the wild west in terms of privacy. Past work has demonstrated the feasibility of skill-squatting.

The paper is the first to explore the extent to which privacy-violating skills that are intentionally authored as such were violated.

===== Weaknesses =====

The paper is not systematic in terms of either its claims or its results:

- The paper claims that it is the first to characterize the security threats of the Amazon Alexa platform. This seems untrue, given the past work on skill squatting, for example.

- The privacy-violating nature of Alexa Skills was based largely on reviews from the Alexa skills store, with a smaller number of skills directly tested.

===== Detailed comments for the author(s) =====

This paper covers an important topic in a relatively under-explored area. It also has some important findings, specifically that Alexa skills often violate user privacy (according to reviews).

The paper meanders, however, into tangents such as legal analysis (e.g., on COPPA) where the authors are unqualified to offer legal opinions. And much of the paper deals with an unsystematic evaluation of reviews in the Alexa skills store. Many of those reviews may not be valid (how often do you trust reviews you read on Amazon that refer to the shipping process, as opposed to the product, for example!), either due to bias, uninformed opinion, or other factors. Table 3 provides some examples of this: "Asks for you to buy additional sounds" means nothing; "scared the kid" may be more of a commentary on the kid than on the skill.

The section on mitigation in similarly non-technical and uninformative. The syggestions are trite (e.g. "ensure description to function fidelity", which essentially says nothing more than making code match spec... OK, but that's a hard problem that software engineers have been studying for decades).

Overall, this kind of analysis on skill reviews needs to be much more systematic. The authors could look into inductive coding, for example, as a means of performing text analysis. Or text-based clustering, if the authors prefer a more quantitative approach.

This is a good problem area. The authors should be more rigorous in both the text analysis of the reviews, as well as in the dynamic testing; for an example of dynamic testing in this area, the authors might look at recent work from Northeastern, or the IoT Inspector project, both of which have done similar testing of IoT devices (including Alexa), as far as third-party data leaks.