Title:

How does variation in AI performance affect trust in AI-infused systems: A case study with in-vehicle voice control systems

Objectives:

With the development of artificial intelligence (AI) and the computing power in personal devices (such as smartphones, smart home systems, and smart cabins in vehicles), AI-infused systems are becoming increasingly popular. At the same time, commercial companies can frequently update AI-infused systems over the air (OTA), which raises concerns about the unstable performance of the AI-infused systems due to uncertainty and inexplicability characteristics of AIs. Specifically, although the OTA aims to improve the overall performance of the system, it may downgrade the performance of certain functions temporarily to balance the computing power or simply due to the instability of new algorithms.

This variation in the system performance may impact users' trust in the system and further affect their reliance on the system. It has been widely acknowledged that trust in a system is a dynamic process. For example, previous research has pointed out that experiencing system failure can undermine users' trust in the system and it takes a relatively long time for them to rebuild their trust. Thus, inappropriate strategies in adjusting the system performance may cause users to stop using the systems. However, previous research on dynamic trust usually targeted systems with rare failures or systems that are safety-critical, such as driving automation. For systems that have relatively high failure rates but low-risk outcomes (such as voice control systems), users' trust in the system might be based on their interactions with the systems over a relatively long period and hence, it is interesting to investigate: 1) the relationship between users' perceived system reliability (as measured by perceived successful rate) and their trust in a system; 2) and whether and how system performance variation can affect users' trust in the system.

To answer these research questions, in this study, we designed a Wizard of Oz (WoZ) system to simulate voice control systems (VCSs) in smart cabins. We chose VCS in the study as it is common in daily life, frequently updated in daily devices (such as smartphones and smart cabins), and with relatively unstable performance. The response accuracies of the VCS varied throughout the study to simulate OTA. The findings of this study can help AI-infused system providers (such as designers of smart cabins) select better roadmaps to update low-risk AI-infused systems if certain functions must be compromised and design strategies to re-attract users if negative events have happened.

Approach:

In total, 27 participants (17 male and 10 female) with an average age of 28 years old (min: 16, max: 54, standard deviation: 10) completed the study. To simulate the VCS, participants were told that they were recruited to test a VCS system developed by our research team. The simulated VCS displayed a virtual animation after being called and provided audio feedback (Figure 2 in supplemental materials). The pre-defined questions were related to the vehicle functions (e.g., opening the window) or infotainment systems (e.g., playing music) and the participants could easily tell the correctness of the responses.

Upon arrival, the participants signed the consent and were instructed on how to use the VCS. Each participant was required to give three batches of 10 queries to the VCS. Between batches, the participants were told that the system version had been upgraded. To simulate the variation in VCS performance, three actual correct rates (ACR) of responses were used, i.e., 50%, 70%, and 90%, leading to 27 possible combinations in the three batches (3*3*3). The failures happened randomly in each batch, but the participants were not informed of the ACR of the VCS. Participants were randomly assigned to one of the 27 combinations.

After participants finished each batch, they completed a *Trust between People and Automation* questionnaire and provided their perceived correct rate (PCR) of the VCS version they had just experienced. The trust score ranges from 1 ("not at al"l) to 7 ("extremely"). No participant discovered that the VCS was fake.

Findings:

First, as expected, users' PCR of the system can be influenced by the ACR of the current system (F(1,50)=4.05, *p* = .049); but not by the ACR in previous versions (*p*=.2), as demonstrated by Model 1 and Figure 3 in the supplemental material.

Then, we explored the influential factors of users' trust in the VCS. To account for the individual differences in users' propensity to trust, we calibrated one's trust in the VCS based on his/her initial trust in the VCS reported before the experiment.

- Participants' trust in the systems was positively associated with the PCR of the current VCS version (Model 2, F(1,42.6)=4.22, *p*=.046). However, the change in PCR between the current and previous VCS versions did not affect users' trust in the current version (*p*=.4).
- Instead, as shown in Model 3, we observed an interaction effect between users' perceived change of correct rates (between the last version and the current version, or "Change in PCR" for short) and the users' perceived correct rate of the last version (i.e., PCR of last version) (F(1,34.5)=7.46, *p*=.01). We found that

with the increase of the PCR of last version, the influence of the Change in PCR on users’ trust in VCS reduced.

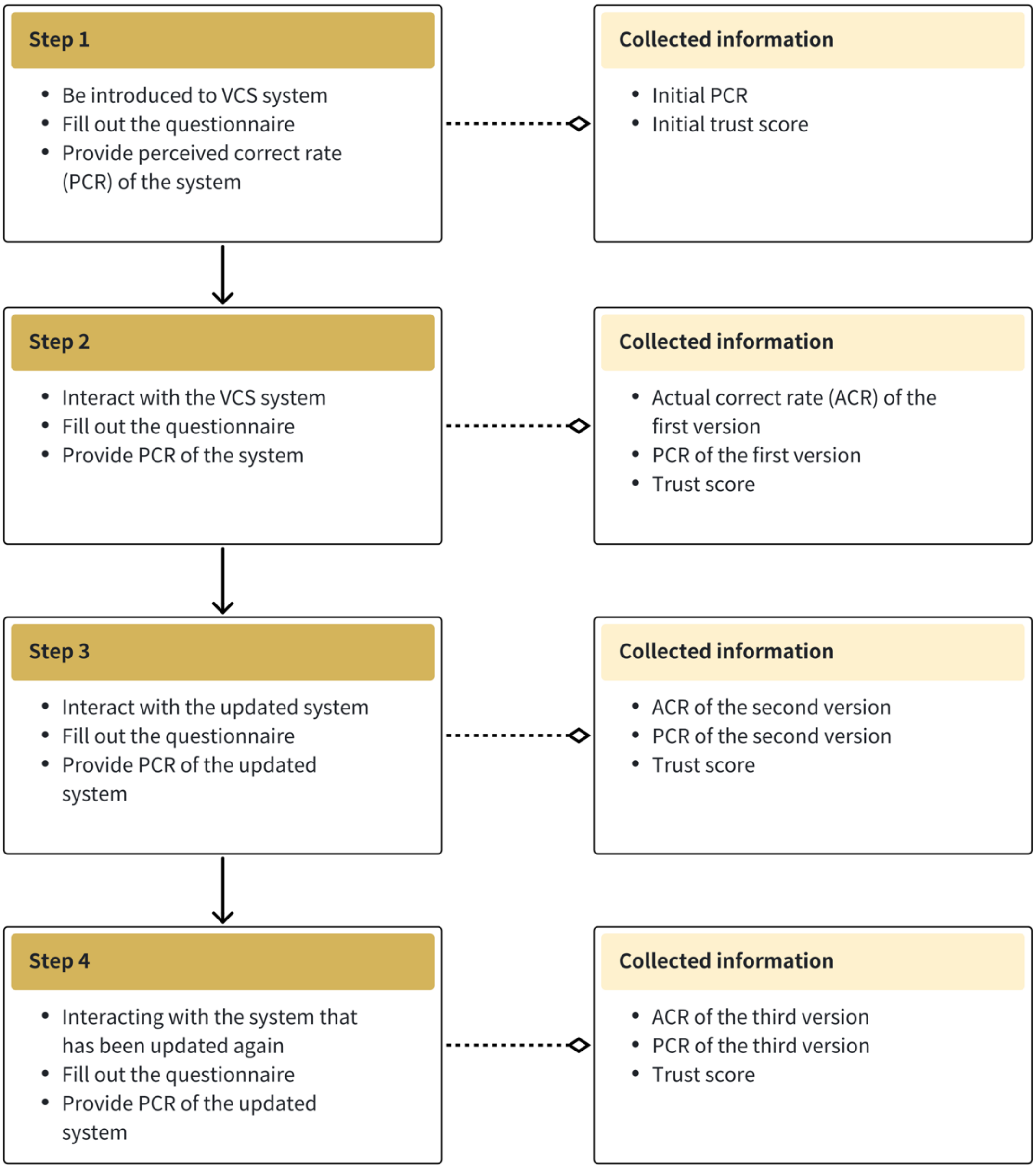
- The perceived pattern of the system upgrades (i.e., perceived correct rates of the first two versions of the VCS), surprisingly, did not affect the trust in the 3rd version (Model 4, $p>.05$), potentially because users’ impression of the first two versions fades out over time.

Takeaways:

- This research contributes to the growing body of knowledge on the relationship between AI evolution and users’ trust.
- The way the system evolves can impact users’ trust in the current version of the system, but this effect fades out with time. Specifically, for the VCS system we explored, only the last version of the system can affect users’ trust in the current system version, but the 1st version would not affect users’ trust in the 3rd version.
- The findings also indicate that users’ perception of the system performance, instead of the actual performance of the system had a stronger influence on users’ trust in the system. Efforts can be made to increase the users’ awareness of the system improvement if we aim to increase users’ trust in a system.
- With higher system performance, the marginal benefits of the system improvement decrease. The system designers may need to consider this effect to balance the cost and return of certain system optimization.

Supplemental Material

- **Experimental information**



Step 1

- Be introduced to VCS system
- Fill out the questionnaire
- Provide perceived correct rate (PCR) of the system

Collected information

- Initial PCR
- Initial trust score

Step 2

- Interact with the VCS system
- Fill out the questionnaire
- Provide PCR of the system

Collected information

- Actual correct rate (ACR) of the first version
- PCR of the first version
- Trust score

Step 3

- Interact with the updated system
- Fill out the questionnaire
- Provide PCR of the updated system

Collected information

- ACR of the second version
- PCR of the second version
- Trust score

Step 4

- Interacting with the system that has been updated again
- Fill out the questionnaire
- Provide PCR of the updated system

Collected information

- ACR of the third version
- PCR of the third version
- Trust score

Figure 1. Flow diagram of experiment process.

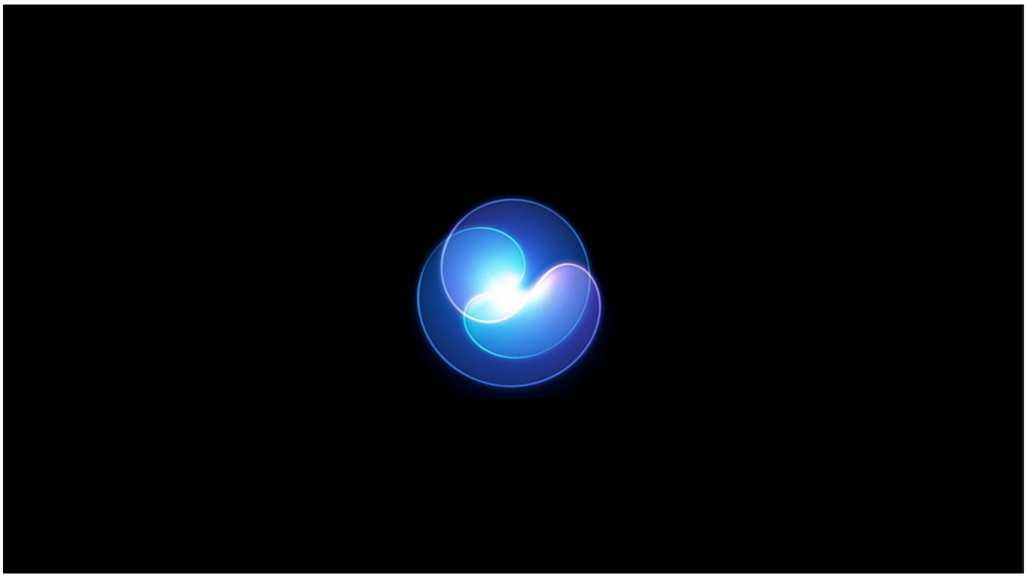


Figure 2. Screenshot of the VCS virtual animation.

- **Statistical Results**

Table 1. Statistical results for the metrics.

DVs	*IVs*	*F-value*	*p-value*
Model 1: PCR	ACR	F (1,50) = 4.05	.049
	Change in ACR	F (1,50) = 1.43	.2
	ACR * Change in ACR	F (1,50) = 1.15	.3
Model 2: Trust in the 2nd and 3rd version	ACR	F (1,42.6) = 4.22	.046
	Change in ACR	F (1,31.9) = 0.79	.4
	ACR * Change in ACR	F (1,31.3) = 0.94	.3
Model 3: Trust in the 2nd and 3rd version	PCR of last version	F (1,39.2) = 5.14	.03
	Change in PCR	F (1,37.1) = 1.29	.3
	PCR of last version * Change in PCR	F (1,34.5) = 7.46	.01
Model 4: Trust in the 3rd version	PCR	F (1,17) = 0.59	.5
	PP	F (4,17) = 0.28	.9
	PP * PCR	F (4,17) = 0.21	.9

Note: DVs stands for dependent variables; IVs stands for independent variables; PCR stands for perceived correct rate; Change in PCR = PCR of current version – PCR of last version; ACR stands for actual correct rate; PP stands for perceived pattern of the system upgrades. In this table and the following figures, trust is calibrated by the initial trust, i.e., trust = trust in the current version – initial trust.

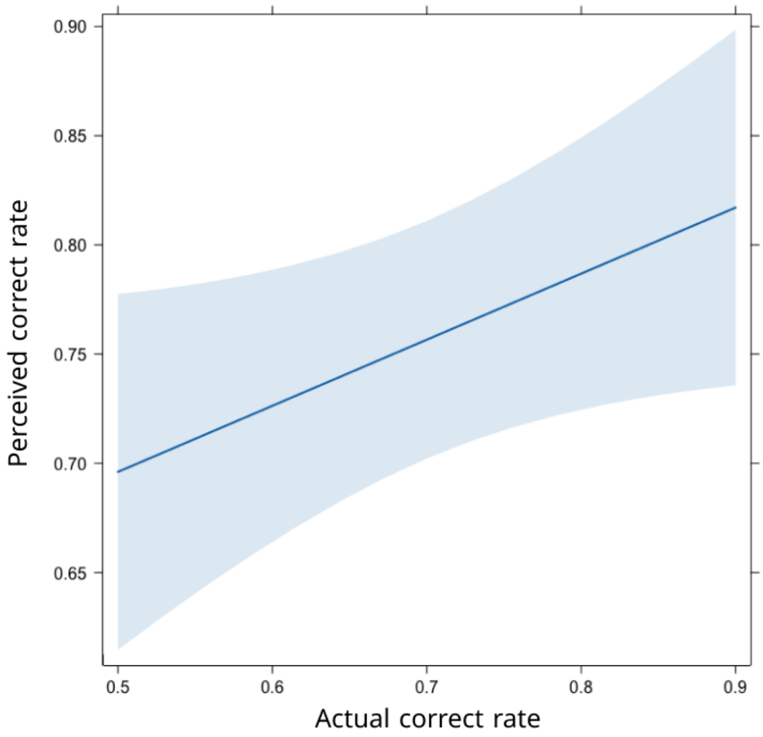


Figure 3. The effect of ACR in model 1. The shadow represents the 95% confidence interval (CI) of PCR.

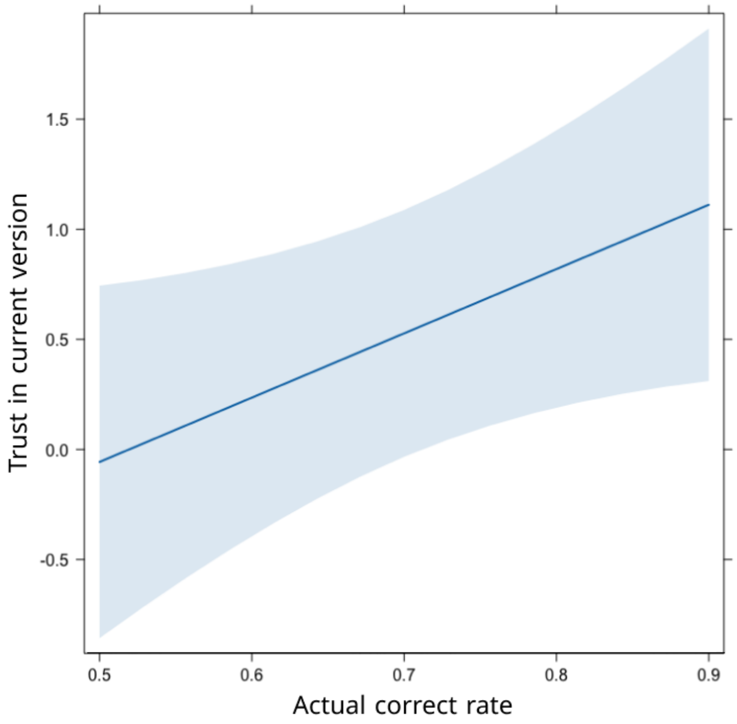


Figure 4. The effect of ACR in model 2. The shadow represents the 95% confidence interval (CI) of trust in current version.

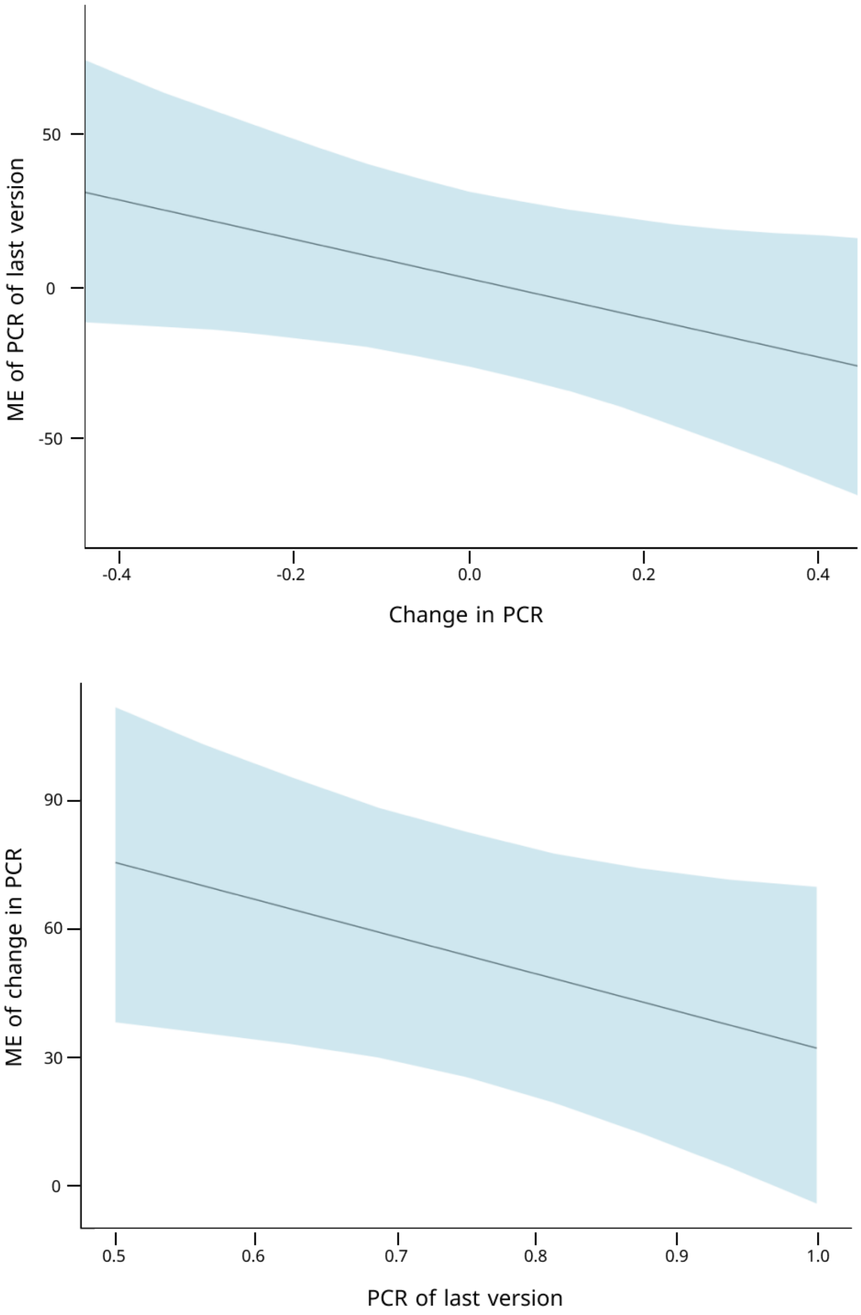


Figure 5. The marginal effects (ME) of the interaction effects between PCR of last version and the change in PCR in model 3. The shadow represents a 95% confidence interval (CI) of the estimated effect.