Dangerous Skills

Understanding and Mitigating Security Risks of Voice-Controlled Third-Party Functions on Virtual Personal Assistant Systems

Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, Feng Qian

Voice Assistants

Alexa, play Today's Hits on Pandora

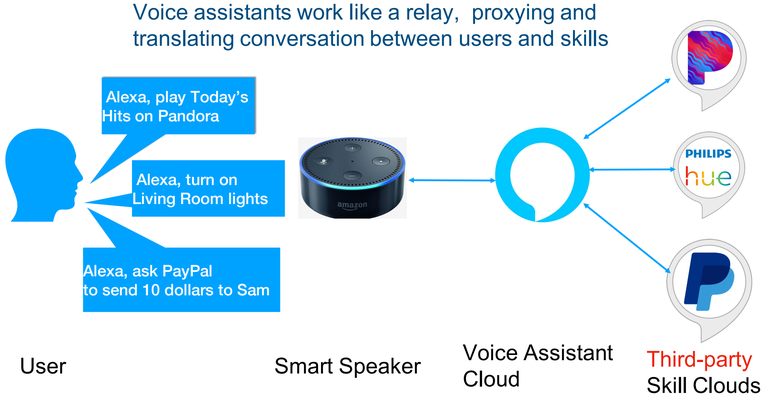
Alexa, turn on Living Room lights

Alexa, ask PayPal to send 10 dollars to Sam

Alexa, ask Medical Assistant to give me my diagnosis

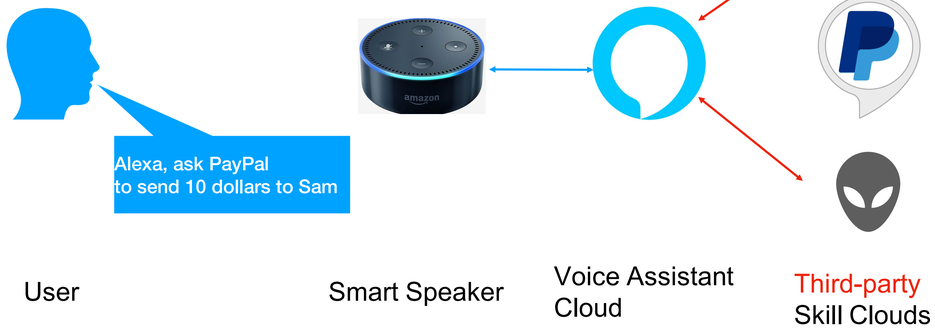
System Mechanism

Voice assistants work like a relay, proxying and translating conversation between users and skills



Voice Squatting Attack

Voice assistants may fail to understand user's intention, and mistakenly invoke wrong skills



Attack Demonstrations

Voice Squatting through invocation name extending

Capital One → Capital One Please / My Capital One / Capital One App

	Alexa	Google
invocation name + "please"	10/10	0/10
"my" + invocation name	7/10	0/10
"the" + invocation name	10/10	0/10
invocation name + "app"	10/10	10/10

Voice Squatting through similar pronunciation

Capital One → Capital Won / Captain One / Capitol One

Alexa		
Amazon TTS	Google TTS	Human
10/17	12/17	> 50%
Google		
Amazon TTS	Google TTS	Human
4/7	2/4	> 50%

Measurements: Error-Prone AIs

Invocation Names → Record → Voice Recordings → Play → Voice Assistant Platforms → Recognition → Helper Skill

- 100 invocation names for each platform
- Human subjects & TTS services
- Those voice assistant platforms are error-prone when recognizing voice commands

	TTS services	Human subjects
Alexa	30%	57%
Google	9%	10%

Recognition Mistake Rates

- Florid state quiz → Florid snake quiz
- Rent Europe → Read your app

Measurements: Conflicting Skills

- 19% (3718) skills： same pronunciation — 66 skills were named as "cat facts", and provided similar functions.
- 2.7% (531) skills: same pronunciation， but different spelling
- 1.8% (345) skills: longest prefix matching
- Interesting cases
  - dog fact → me a dog fact

Both "SCUBA Diving Trivia" and "Soccer Geek", registered "space geek" as invocation names

Defense

UIC: User Intention Classifier

Classify user's intention as context switching or not

SRC: Skill Response Checker

Identify suspicious skill response, such as fake skill recommendation

INDIANA UNIVERSITY

UNIVERSITY of VIRGINIA