<주요 Q&A>

AI가 만드는 Cyber World – 보이는 대로 믿으시겠습니까?

Q1. 딥페이크 + 딥보이스를 감지하고 예방할 수는 없나요?

딥페이크와 딥보이스를 감지하고 예방하는 방법은 크게 탐지 솔루션을 사용하는 사후 대응과, 발생하기 전 사전에 차단하는 방법이 있습니다. 딥페이크 이미지나 영상 관련 사후 대응은 본 발표에서 설명 드린 탐지 솔루션으로 대응 가능하며, 딥페이크 보이스 대응의 경우는 사전 차단을 위해 사용자간 별도의 솔루션 앱을 설치하여 사용자 생체 인증을 통한 사용이 가장 적절할 것으로 생각합니다. (딥보이스의 경우 사후 탐지가 매우 어렵습니다.)

Q2. 딥보이스를 통한 보이스 피싱이 나오면 위험하겠네요.

맞습니다. 이미 19년 해외에서 딥보이스 피해(송금 피해)가 발생했습니다. 시만텍에서는 알려지지 않은 피해 사례가 더 많은 것으로 파악하고 있습니다.

Q3. 딥페이크를 활용한 물리보안 솔루션은 어떤 것이 있는지 궁금합니다.

딥페이크 탐지 기술을 활용한 물리적 보안 솔루션은 딥페이크 탐지만을 위한 솔루션이라기 보다 사용자 인증 솔루션으로 설명드릴 수 있겠습니다. 물리적 보안은 게이트웨이나 보안 카메라, 사용자 인증 시스템 등이 가능한데 사용자의 liveness를 실시간 확인하는 사례를 딥페이크 탐지 엔진과 웹 캠을 결합하여 구현해 볼 수도 있을 것 같습니다.

Q4. 현재 딥페이크 탐지 기술 개발 과정에 있어 아직까지도 해결되기 힘든 문제는 어떤 것들이 있는지 궁금합니다.

딥페이크 탐지는 딥러닝의 GAN 모델에서 비롯된 artifact를 탐지하는 것으로 말씀드릴 수 있는데, 이미지를 추출하는 과정에서 compression이 발생하면 artifact 자체가 사라질 수 있어서 탐지성능이 저하되는 경우가 있습니다.

Q5. 딥페이크 탐지 기술의 정확도(신뢰도)는 현재 어느 정도 수준인지 궁금합니다.

> 자체 테스트 과정에서 실제 얼굴 5만장, 딥페이크 5만장 해서 총 10만장의 사진을 대상으로 테스트한 결과 1장 오류가 났습니다. 즉, 99.999%의 결과를 보였습니다.

Q6. 딥페이크 분석에 어떤 알고리즘을 사용하나요?

> 삼성SDS 팀나인에서 자체 개발한 알고리즘이고, 이미지를 주파수 기반으로 변환하여 특징을 비교 분석하고 있습니다.