창작의 신기원인가, 열린 사회의 적인가?

딥페이크

권오현 계간 스켑틱 편집자

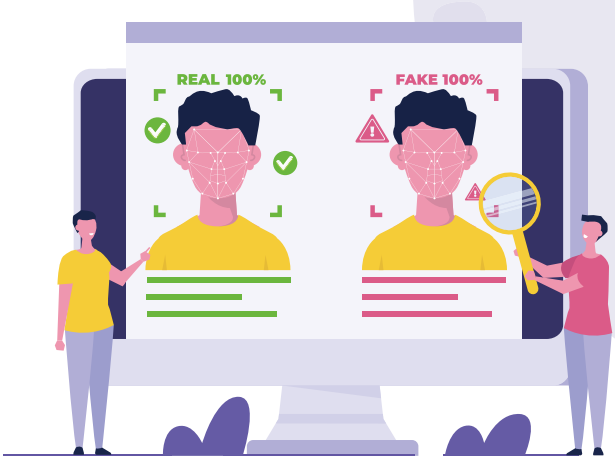
2024년 8월, 남성들이 여성 지인이나 불특정인, 연예인의 얼굴 사진을 음란물과 합성해 텔레그램(Telegram)을 통해 유포하는 딥페이크 디지털 성범죄가 광범위하게 일어나 우리 사회에 큰 충격을 주었다. 무엇보다 우려해야 할 지점은, AI 기술의 발전으로 누구나 쉽게 이런 디지털 성범죄를 죄의식 없이 저지를 수 있게 됐다는 점이다.

이미지 생성 기술로서의 딥페이크

딥페이크는 쉽게 말해 AI 기술인 '딥러닝(Deep learning)'과 '가짜'를 뜻하는 '페이크(Fake)'의 합성어로, AI 기술을 이용해 실제 존재하는 것처럼 보이는 사진, 비디오, 오디오를 모두 가리킨다. 딥페이크라는 단어는 2017년 처음 등장했으며 미국 온라인 커뮤니티 레딧(Reddit)에서 'deepfakes'라는 아이디를 쓰는 회원이 기존 영상에 유명인 얼굴을 합성한 가짜 콘텐츠를 제작, 게시한 데서 유래했다. 이렇게 딥페이크 콘텐츠는 온라인 커뮤니티와 소셜 미디어를 중심으로 퍼져나갔고 딥페이스랩(DeepFaceLab), 페이스스왑(Faceswap) 같은 오픈 소스 형태의 영상 합성 제작 프로그램이 배포돼 대중화됐다.

딥페이크는 2014년 등장한 딥러닝 기술인 'GAN(적대관계생성신경망, Generative Adversarial Networks)'에 기반한다. GAN은 원본과 흡사한 데이터를 만드는 생성 모델과 그 생성 모델에서 원본과 다른 점을 찾아내는 분류 모델을 서로 경쟁시키는 것을 반복 실행해, 원본과 흡사한 데이터를 만들어내는 알고리즘이다.

GAN을 이용한 이미지 생성 기술은 드라마나 영화에서 각광받았다. 배우가 직접 촬영하기 어려운 부분에 대역을 사용한 후 대역의 얼굴을 배우의 얼굴로 변환하는 '페이스에디팅(Face-editing)' 기술, 배우의



얼굴(주름이나 기타 잡티)을 보정하는 '페이스에이징/디에이징(Face-aging/deaging)' 기술은 GAN을 통해 만들어진다. 영화 '어벤저스-엔드게임', '아이리시맨'이 대표적인 사례다.

딥페이크는 우리 사회의 신뢰도와 직결된다

그러나 정교한 가짜 이미지라는 딥페이크 기술의 특징은 범죄나 사기에 악용될 소지가 크다. 2018년 4월 미국 온라인 뉴스 매체 버즈피드(Buzzfeed)가 유튜브(YouTube)에 게재한 한 영상이 화제가 됐다. 버락 오바마(Barack Obama) 전 미국 대통령이 당시 대통령이었던 도널드 트럼프(Donald Trump)에게 욕을 하는 내용이었기 때문이다. 이는 조작 영상이었다. 버즈피드가 딥페이크 기술의 위험성에 대해 경고하기 위해 영화감독 조던 필(Jordan Peele)과 함께 만들어 배포한 것이다.

이에 구글(Google), 메타(Meta), 아마존(Amazon), 오픈AI(OpenAI) 같은 AI 기술 공급 기업은 딥페이크 기술이 일으키는 새로운 범죄에 대응하기 위한 규제 및 표준안을 제정하고 있다. 그 핵심은 첫째, AI 시스템 발표 전 제품의 안전성 보장(보안 테스트 시행). 둘째, 보안을 최우선으로 하는 시스템 구축. 셋째, 대중의 신뢰 확보(AI 생성물임을 표시, 투명성 강화 등)다. 2024년 2월, 구글은 메타, 오픈AI와 함께 콘텐츠 출처에 대한 기술표준인 콘텐츠 인증 정보(Content Credentials)를 수용하기 위해 C2PA(콘텐츠 출처 및 진위성 연합, Content Provenance and Authenticity)에 가입할 것이라고 발표했다 C2PA의 기술로 콘텐츠에 메타 데이터를 삽입하면 콘텐츠의 출처를 확인할 수 있어 해당 콘텐츠가 AI로 제작된 것인지의 여부를 알 수 있다.

국가 차원에서도 딥페이크 범죄에 맞서는 법안을 제출, 시행하고 있다. 2024년 5월, EU(유럽연합, European Union) 이사회는 세계 최초로 AI를 포괄적으로 규제하는 AI법(Artificial Intelligence Act)을 승인했다. AI법은 인간 중심적이고 신뢰할 수 있는 AI의 채택을 촉진하는 동시에, EU 내 AI 시스템의 유해한 영향으로부터 건강·안전·민주주의, 법치 및 환경을 보호하려는 목적을 갖고 있다. AI법은 EU 소재지와 관계없이 EU에서 AI 시스템을 제작, 사용, 수입, 배포하는 모든 사람에게 적용되며, AI를 잠재적 위험에 따라 금지·고위험·제한적 위험·최소한의 위험이라는 네 가지 범주로 분류한다. 이 중 딥페이크는 제한적 위험에 해당한다. 딥페이크는 AI 시스템 공급자(AI 시스템 개발)와 사용자(Deployer: AI 시스템 사용) 모두에게 투명성 의무를 부과한다.

미국에선 최근 들어 주 정부 차원에서 AI 기술의 악용 가능성을 제한하는 법안을 많이 제출, 시행하고 있다. 특히 선거용 가짜 뉴스, 음란물 제작과 유포를 금지하고 그 형량을 늘리고 있다.

우리나라에선 현재 유럽처럼 AI를 전반적으로 포괄하는 법은 없다. 딥페이크를 이용한 범죄에는 대개 '정보통신망 이용촉진 및 정보보호 등에 관한 법률'이 적용된다. 그러나 2024년 5월 과학기술정보통신부가 관계부처 합동으로 마련한 '새로운 디지털 질서 정립 추진계획'에는, 디지털 심화 쟁점 해결을 위한 20대 정책과제가 있으며, 이 중 8대 핵심 과제에 '딥페이크를 활용한 가짜뉴스 대응'이 포함돼 있다.

기술이 줄 장밋빛 미래만 기대하고 모든 것을 자율로 맡기는 것은 디스토피아나 다름없다. 규제는 악이 아니라 필수다. 우리의 미래가 진정으로 밝으려면 무너져가는 신뢰를 다시 회복하는 것이 선결돼야 한다. 딥페이크 규제는 그 상징적인 사건이다.