딥페이크 기술의 발전과 저널리즘의 새로운 위협
– 이미지공정성 확보를 위한 영상 팩트체크 필요성

박주일 MBC 보도본부 뉴스콘텐츠 취재2부 영상기자

01 논의를 시작하며

인류 역사 이래로, 인간의 지각능력에서 가장 우위에 있었던 것은 주로 '시각'이었다. 동·서양을 막론하고 'Seeing is believing'이나 '백문이 불여일견(百聞不如一見)'과 같은 속담들이 통용되었던 이유도 바로 이처럼 시각 인지(認知)에 대한 확고한 믿음 때문이었다. 눈으로 보는 사물과 현상의 이미지는 가장 강력한 진실의 척도였고, 영화나 드라마에 단골로 등장하는 대사조차 '내 두 눈으로 똑똑히 보기 전까지는 절대 믿을 수 없다'였음을 보면, 무언가를 판단함에 있어 시각에 대한 의지는 불가침의 영역이나 다름없었다.

[그림 1] 오바마 전 미국대통령 딥페이크 영상

하지만 AI와 딥러닝으로 대표되는 디지털 기술의 발전과 이로 인한 미디어 제작기법의 진화는 유사 이래 도전받지 않았던 시각 인지에 대한 인류의 오랜 믿음을 송두리째 흔들고 있으며 그 정점에 서 있다 할 수 있는 것이 최근 들어 사회문제로도 대두되고 있는 '딥페이크' 영상이다.

쉽게 말하자면 '가짜 동영상'의 한 부류인 딥페이크는 고도화된 이미지 합성 기술의 산물인데, 지난 2018년 공개된 오바마 전 미국 대통령의 거침없는 발언 영상[1]이 전 세계적으로 화제가 되면서 딥페이크 영상은 저널리즘 영역에서도 중요한 논란거리로 떠올랐다. 영상에서 오바마는 귀를 의심케 하는 비속어를 사용해 트럼프 미국 대통령을 모욕하는 발언[2]을 했으나 사실 이는 배우이자 영화감독인 조던 필(Jordan Haworth Peele)의 성대모사에 딥페이크 기법으로 제작한 오바마의 백악관 영상을 덧입힌 것이었다. 필 감독과 함께 이 영상을 제작한 버즈피드는 딥페이크 영상의 위험성과 사실 구분의 어려움을 보여주기 위해 이러한 기획을 시도했다고 밝혔는데, 결과적으로 대중들에게 현재의 딥페이크 기술이 얼마나 사실 왜곡의 위험성을 가졌는지를 강렬하게 인지시켜 주었다.

딥페이크와 관련된 논란은 국내에서도 어렵지 않게 찾아볼 수 있다. 지난 4월 대략 100여 명이 넘는 K-팝 스타들의 딥페이크 음란물 영상이 유통된 인터넷 사이

1) BuzzFeedVideo (2018. 4. 17). 'You Won't Believe What Obama Says In This Video!'. URL: https://www.youtube.com/watch?v=cQ54GDm1eL0
2) 앞의 BuzzFeedVideo (2018. 4. 17). "President Trump is a total and complete dipshit(트럼프 대통령은 완전히 쓸모없는 인간이야)".

트에 대해 경찰이 수사에 착수했는데, 중남미 등 해외에 서버를 둔 이들 업체가 유통한 동영상의 숫자가 3천 건이 넘을 정도로 방대해 이를 내려받은 사용자들에 의한 2차, 3차 확산까지 우려되었다.[3] 또한 네덜란드의 사이버보안 기업인 딥트레이스(Deeptrace)의 보고서에 따르면 인터넷 공간에 떠돌고 있는 딥페이크 포르노 영상이 두 배로 확산하는데 약 9개월밖에 걸리지 않으며 그 중 약 4분의 1이 K-팝 여자 아이돌 그룹과 관련된 영상[4]이라고 하니 한국에서의 상황은 더욱 심각하다 할 수 있다.

이처럼 갈수록 고도화되어가고 있는 딥페이크 영상의 부작용은 단순한 화젯거리의 차원을 훌쩍 넘어섰으며 저널리즘 영역에서도 시급한 위협으로 현실화하고 있다. 상술한 사례들에서 알 수 있듯 딥페이크 기법의 정교함은 이제 개인이 판단하기 어려운 수준인 데다 특히 SNS가 주요 미디어 플랫폼으로 자리 잡은 최근의 저널리즘 특성상 사실 확인이 이뤄지지 않은 이미지와 영상이 순식간에 확산될 가능성은 언제든 상존한다. 때문에 전통적인 언론의 팩트체킹 노력은 이제는 그 범위를 확장해야 하며 왜곡되지 않고 신뢰할 수 있는 이미지를 전달하려는 '이미지 공정성' 확보가 언론의 새로운 과제로 떠올랐다.

02 딥페이크의 기술적 배경과 진화 형태

가. 딥페이크의 개념

사전적 의미에서 딥페이크(Deepfake)는 기계학습(machine learning) 기법의 하나인 '딥러닝(deep learning; 심층학습)'과 가짜를 의미하는 '페이크(fake)'가 결합된 합성어로 인공지능을 기반으로 원본 이미지 위에 다른 이미지를 중첩하거나 결합시켜 원본과 쉽게 구분할 수 없는 가공의 이미지와 소리를 갖도록 만들어지는 영상 혹은 그 제작과정을 말한다.[5] 지난 2017년, 미국의 온라인커뮤니티인 레딧(Reddit)의 'Deepfakes'라는 계정에 유명 헐리웃 스타들의 얼굴을 조작한 가짜 영상들이 업로드되면서 딥페이크라는 용어가 최초로 사용된 것으로 알려져 있는데, 이후 레딧은

3) 조희형 (2020. 4. 20). '연예인 1백여명 얼굴이… 딥페이크 누가 만드나'. 〈MBC〉. URL: https://imnews.imbc.com/replay/2020/nwdesk/article/5739058_32524.html

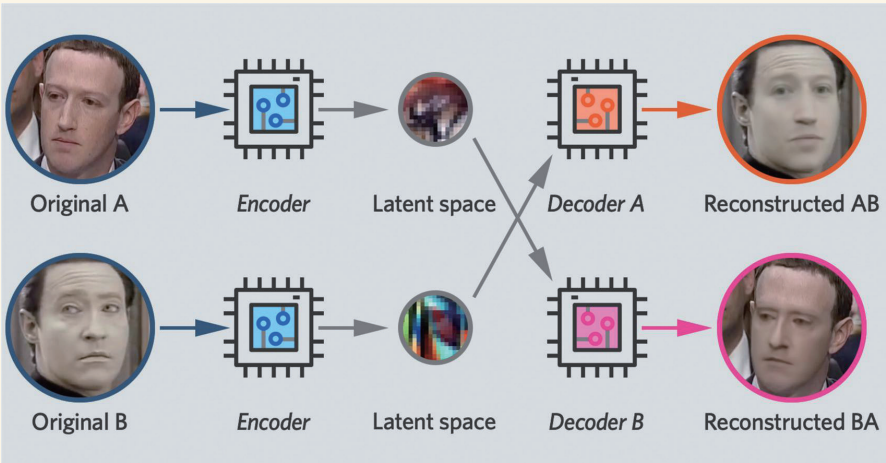
4) Rory Cellan-Jones (2019. 10. 7). Deepfake videos, doubles in nine month' 〈BBC News〉. URL: https://www.bbc.com/news/technology-49961089

5) Marie-Helen Maras & Alex Alexandrou (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. 〈Int'l Jounal of Evidence and Proof〉, Vol.23 p.255~256, 참조 ; 오세욱·이소은·최순욱 (2017). 기계와 인간은 커뮤니케이션할 수 있는가?: 기계학습을 통해 본 쟁점과 대안. 〈정보사회와 미디어〉, 제18권 제3호, p.63~96.

딥페이크와 관련된 서브레딧(특정분야를 모아놓은 카테고리)을 종료했으며 사용자 'Deepfakes'는 여기에 사용된 소스코드를 모두 공개했다.

이처럼 주로 얼굴 합성을 대상으로 하는 딥페이크는 이미지를 추출·학습·생성하면서 가상 이미지를 만들어내는데 최근 들어 '생성적 적대 신경망(GAN;Generative Adversarial Network)' 방식의 기계학습기술을 사용하며 더욱 정교하게 진화하고 있다. 이는 최대한 원본과 같은 이미지를 만들어 구분이 어렵도록 하는 인공지능 네트워크와 반대로 최대한 원본과 다른 점을 찾아내어 구분하려는 인공지능 네트워크가 서로 경쟁적으로 결과물을 주고받으며 최종적으로는 가장 진짜 같은 가짜 이미지를 완성하는 방식이다.[6]

[그림 2] 딥페이크 얼굴변환 방식 개념도[7]



이렇게 만들어진 딥페이크 영상이 위협적인 까닭은 딥러닝을 반복하는 인공지능 기술의 특성과도 맞닿아 있다. 지난 2016년 인간과 기계의 대결로 대중들의 지대한 관심을 받았던 구글의 알파고와 바둑기사 이세돌의 바둑대국[8]에서 많이 알려졌다시

6) 박준·조영호 (2019). '딥페이크 영상 탐지 관련 기술 동향 연구', 〈한국소프트웨어종합학술대회 논문집〉, p.725. 참조

7) Timothy B. Lee (2019. 12. 16). I created my own deepfake—it took two weeks and cost $552. 〈ars technica〉. URL: https://arstechnica.com/science/2019/12/how-i-created-a-deepfake-of-mark-zuckerberg-and-star-treks-data/3/

8) 총 5차례 열린 대국에서 알파고는 이세돌을 상대로 1패만을 내어주고 나머지 4게임은 승리하였다. 유튜브를 통해 생중계된 이 대국은 세계적으로 8천만 명이 넘는 시청자가 관람하였다.

피, 빅데이터를 기반으로 한 인공지능 프로그램의 진화는 초기의 학습 단계만 거치고 나면 점점 강해지는 자신과의 게임을 반복하면서 스스로 빅데이터를 구축하는 경지에 이르렀다. 같은 맥락에서 딥페이크로 조작된 영상의 정교함 역시 시간이 지날수록 차원이 다른 완성도로 구현될 것이 분명하다.

나. 딥페이크 제작의 보편화–자동화와 앱(App) 기반 제작 환경

앞서 언급된 다양한 기술적 배경과 사용되는 연산능력 등을 고려했을 때, 딥페이크 이미지가 고도화되는 것은 사실이지만 여전히 전문적인 장비나 설비가 있어야만 제작 가능한, 일종의 전문적 영역이라는 판단을 내릴 수도 있다. 불과 5~6년 전만 하더라도 정밀한 딥페이크 영상 제작에는 고급 GPU(그래픽 작업을 총괄하는 연산장치)가 장착된 여러 대의 컴퓨터가 필요했으며 그 때문에 주로 영화제작의 CG 작업에 활용되거나 특수효과 및 VFX 제작 등에 한정적으로 활용되었다. 이는 회화에서 사진 그리고 영상으로 이어지는 이미지 제작 기술의 발전 단계마다 일관되게 유지되었던 특징인데 고도의 전문가 집단만이 가능한 이미지 처리영역이 존재했고 그렇기에 대중은 이렇게 가공된 영상이 활용되는 범위나 방식을 의심하지 않아도 되었다.

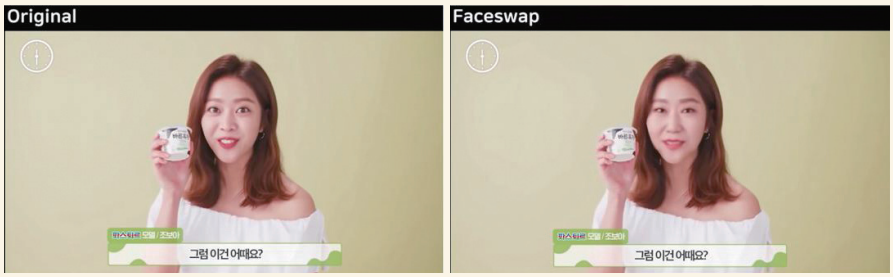
하지만 이러한 기존의 이미지 조작방식과 달리 최근의 딥페이크 기법이 가지는 주요한 차이점은 이미지 조작에 대한 결정과 기능을 수행하는 전 과정이 모두 진정한 의미에서 가내화(家內化, domestication), 보편화되었다는 점이다.[9] 가정용 컴퓨터의 성능만으로도 충분히 정교한 품질의 영상 합성이 가능해졌고, 프리미어나 포토샵 같은 이미지합성 프로그램의 숙련도와 무관하게 합성할 사진을 고르기만 하면 최적의 결과물이 자동으로 도출되는 소프트웨어[10]들이 등장하면서 더더욱 딥페이크 제작의 장벽은 낮아졌다. 당장 스마트폰 기반의 앱마켓에만 들어가도 수십 가지 종류의 딥페이크 앱들을 무료이거나 소액만으로 설치할 수 있으며 이렇게 간편하게 제작된 영상들조차도 얼핏 봐서는 이질감이 안 느껴질 정도로 딥페이크 기술력은 나날이 발전해가고 있다.

종합하자면, 인공지능과 딥러닝, 앱 기반 소프트웨어로 이어지는 합성 이미지 제

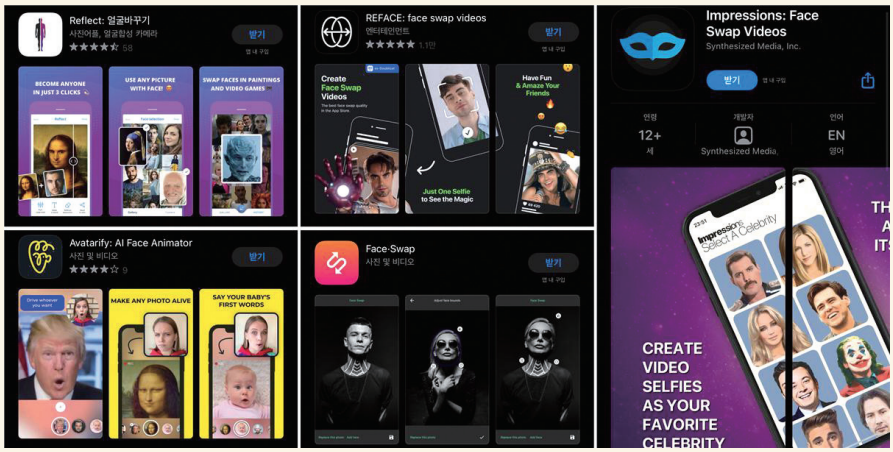
9) 딥페이크는 사람이 사진이나 동영상을 모아서 입력해주기만 하면 잘 조작된 결과물이 산출되기 때문에 영상제작자의 능력에 구애받을 필요가 없다. 게다가 인간은 영상조작에 필요한 사진만 입력할 뿐 조작역할은 수행하지 않으며, 결과물의 적합성만 판단하면 된다. 최순욱·오세욱·이소은 (2019). 딥페이크의 이미지조작 : 심층적 자동화에 따른 사실의 위기와 푼크툼의 생성. 〈미디어, 젠더&문화〉, 제34권 제33호, p.356. 참조

10) 현재는 폐쇄되었으나 초창기 페이크앱(FakeApp)을 위시로, 페이스스왑(FaceSwap)이나 디페이커(Dfaker) 등과 같은 프로그램들이 제공되고 있는 상태.

[그림 3] 페이스스왑(Faceswap)으로 합성한 배우 조보아/라미란 딥페이크 영상[11]



[그림 4] 간단한 조작으로 얼굴합성을 가능하게 하는 딥페이크 어플들



작 프로세스는 더 이상 장비·기술 한계로 인한 딥페이크 제작 장벽이 존재하지 않음을 의미한다. 이제 남은 것은 '누구의 얼굴을 바꿀 것인가'에 대한 1차 선택과 바뀐 얼굴에는 또 '누구의 얼굴을 사용할 것인가' 하는 2차 선택의 문제이다. 기존의 이미지 조작이 '얼마나 사실적으로 보이게 하는가'에 방점이 찍혀있다면 이제는 '얼마나 의외의 인물 조합과 상황을 연출하는가'로 무게중심이 옮겨진 상황이다. 다시 말해, 이제 딥페이크 이슈와 관련한 논의는 기술 영역만큼이나 윤리 영역의 고민도 중요해진 시점이다.

11) [딥페이크] 조보아를 라미란으로. 참고 URL: https://youtu.be/Jd4K47upz0M

03 '가짜 뉴스' 시대의 신성장동력, 딥페이크

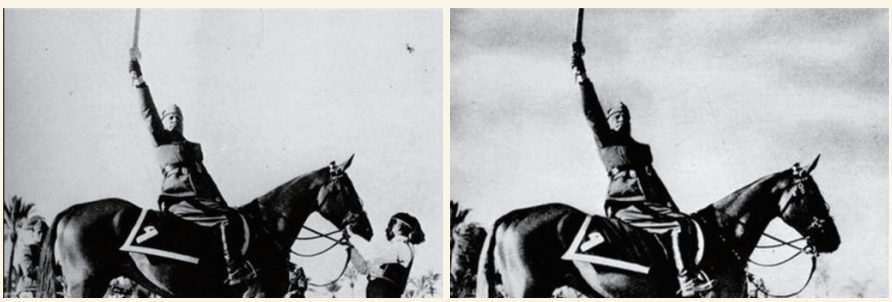
가. 이미지 기반 가짜 뉴스의 태동

텍스트를 기반으로 한 '가짜뉴스'는 지난 2016년 미 대선 과정에서 공화당 후보 트럼프와 그의 캠프가 'Fake news'를 자주 언급한 것이 기폭제가 되어 대중들에게도 익숙한 개념이 되었다. 하지만 개념 보편화 이전에도 가짜뉴스는 문자의 태동과 언론의 역사 이래 줄곧 광범위하게 존재해왔다. 상반되는 개념인 '진짜 뉴스'가 '진실을 다루는 뉴스'인지 '뉴스다운 뉴스'인지 해석차는 있지만 '가짜뉴스' 개념은 대체로 이견 없이 '그릇된 정보를 전달하는 뉴스'로 받아들여지고 있다. 단어와 문장의 조합으로 내용을 구성하는 텍스트의 특성상 창작자인 작가가 불순한 의도만 갖는다면 누구든 손쉽게 가짜뉴스를 생산해 낼 수 있었고, 이는 활자뿐만 아니라 구전(口傳)과 같은 방식으로도 전파될 수 있었다.

반면, 이미지를 기반으로 한 가짜뉴스는 상대적으로 나중에 등장했다고 볼 수 있을 것이다. 물론 회화 시대에도 이미지를 왜곡하는 여러 가지 방법이 있었지만 이는 주로 종교적인 효과를 극대화하거나 예술적 기법의 하나로 이해되었기에 가짜뉴스와는 차이가 있었다.

본격적인 '이미지 가짜뉴스'에 해당하는 조작은 지면 매체인 신문의 등장과 페니 프레스(penny press) 시대[12]를 거치면서 사진을 통해 주로 이루어졌다. 이렇게 왜곡된 이미지 보도는 정치인의 프로파간다를 확산하는 데 중요한 역할을 담당했으며 때로는 텍스트보다 훨씬 강한 파급력과 전달력으로 대중들에게 거짓된 믿음을 설파하였다.

[그림 5] 조련사 도움없이 말에 올라탄 무솔리니[13]



12) 1센트의 소액으로 신문구매가 가능해지면서 노동자계층까지 언론을 소비하게 되었고, 이에 따라 본격적인 지식의 향유가 이루어진 19세기 후반.
13) 1942년 촬영된 이 사진에서, 검을 높이 치켜든 무솔리니의 우상화를 위해 말 조련사에게 도움받는 모습을 삭제했다.

[그림 6] 오바마와 중동 지도자들을 이끄는 무바라크 이집트 대통령[14]

90년대 이후엔 포토샵과 같은 이미지편집프로그램이 등장하면서, 단순히 대상을 삭제하는 수준을 넘어 이미지 합성을 통한 인물의 재배치나 공간 재구성이 가능해지며 좀 더 적극적인 가짜뉴스 생산으로 이어졌다. 이렇게 조작된 사진에는 그럴싸한 사진 설명과 그래픽 등이 결합되었고 대중들은 더더욱 이미지의 조작 사실을 눈치채기가 힘들어졌다.

그런데 최근 들어서는 이러한 이미지 왜곡과 성격이 전혀 다른 가짜뉴스가 등장했는데, 기존의 이미지 조작이 편집자의 적극적인 의지에 따른 것이라면 새로운 형태의 가짜뉴스는 편집자의 의도와 무관하게 행해짐을 특징으로 한다. 즉, 제3자에 의해 생산된 가짜뉴스가 검증 없이 전파되면서 피해를 발생시킨 것인데 국내의 사례로는 최근 빈번하게 발생한 방송사들의 '일베(일간베스트저장소)'[15] 이미지 사용이 대표적이다.

[그림 7] 일베 사이트에서 합성된 노대통령 조작이미지들

14) 이집트의 관영신문 알 아흐람은, 2010년 백악관에서 열린 중동평화협상 중 원래 맨 뒤에 있던 자국 대통령 무바라크의 위치를 바꿔, 오바마 미 대통령과 네타냐후 이스라엘 총리, 압바스 팔레스타인 자치정부 수반 등을 맨 앞에서 이끌며 회의장으로 들어가는 것처럼 조작했다.

15) 극우 성향의 커뮤니티 사이트

방송통신심의위원회의 2019년 자료에 따르면, 2013년부터 2019년 6월까지 일베와 관련된 이미지 사용으로 징계처분을 받은 언론사가 무려 11곳이며 모두 39건의 징계 조치(의견제시 1건, 권고 24건, 주의 11건, 경고 3건)가 내려졌는데, 대부분 노무현 전 대통령 사진을 조작한 일베 사이트의 이미지를 검증 없이 재사용하다가 생긴 일이다. 여기에는 신뢰도를 생명으로 하는 보도프로그램도 다수 포함되는데, 해당 프로그램 담당자는 사과문을 통해 조작된 가짜 이미지가 사용되게 된 배경을 설명했지만 이미 언론의 신뢰도에 커다란 흠집을 남긴 뒤였다.

이 일련의 사건들에서 주목할 점은 합성된 이미지가 가지는 가짜뉴스로서의 전파력과 구분을 어렵게 하는 '인지 난해(recognition difficulty)'에 있다. 짜깁기된 사진과 이미지들은 워낙 교묘하고 정교한 데다 다양한 통로로 퍼져있어, 무의식 중에 가짜뉴스를 전파하는데 가담하는 상황이 발생하는 것이다. 이렇게 전달된 가짜뉴스는 또다시 각종 포털과 커뮤니티의 게시글로 확대·재생산되며 결국 잘못된 여론 형성으로 이어지게 된다. 물론, 이미지가 이곳저곳 전파되는 과정에서 수많은 판독의 기회가 존재했겠지만, '사진은 거짓말하지 않는다'는 말에서도 알 수 있듯 대중들에게 이미지 조작은 여전히 생소한 영역이었고 결국 별다른 의심 없이 무차별적으로 전파될 수 있었다.

나. 팩트체크의 사각 지대, 딥페이크

영상으로 제작된 딥페이크 이미지는 대중들의 이러한 허점을 더욱 강력하게 파고들었다. 정지된 이미지인 사진보다 움직이는 동영상 이미지의 조작 가능성이 훨씬 느슨하게 검증되는 경우가 많은데, 이를 잘 보여주는 사례가 지난 2008년 공개된 BBC의 'Flying penguins' 다큐멘터리[16]이다.

예고편 성격으로 공개된 1분 30초 분량의 이 영상에서 BBC 다큐멘터리 팀은 날지 못하는 새로 알려진 펭귄 중 남극에 존재하는 일부 종이 추운 겨울을 나기 위해 남아메리카 밀림으로 날아간다는 사실을 발견했으며 수천 마리의 펭귄 떼가 빙하 사이로 날아가는 경이로운 모습을 포착해 공개했다. 생물학사에 한 획을 그을 뻔한 이 발견은 아쉽게도 BBC 특수효과팀이 준비한 만우절 맞이 가짜 영상이었는데, 영상을 본 많은 이들이 동물원에 문의 전화를 하고 교과서 수정을 요청하는 등 한동안 해프닝이 이어졌다. '펭귄이 하늘을 난다'는 내용만 놓고 보면 터무니없다는 것이 금방 인

16) BBC (2008). Flying Penguins | World Penguin Day | 〈BBC〉. URL: https://youtu.be/9dfWzp7rYR4

[그림 8] BBC가 공개한 'Flying penguins(하늘을 나는 펭귄)' 영상의 한 장면

지되지만 그것이 정교한 영상과 결합되어 보여졌을 때, 기본적인 상식의 범위가 흔들리기도 한 것이다.

딥페이크 영상이 위협적인 것도 바로 이 부분이다. 하늘을 나는 펭귄보다 훨씬 가능성 있는 상황이나 현실적인 인물의 행동이 딥페이크 기법으로 제작되었을 때 그리고 그 조작 사실이 감추어졌을 때 영상은 혼돈을 야기하거나 최소한 사실로 받아들여질 확률이 굉장히 높다고 할 수 있다.

다. SNS 시대 딥페이크의 위험성

각종 소셜미디어와 온라인 플랫폼의 거대한 팽창은 SNS 시대 딥페이크 영상물에 대한 팩트체킹을 더욱 어렵게 한다. 전통적인 저널리즘의 뉴스 생산 프로세스는 기자–데스크–편집자로 이어지는 2, 3중의 팩트체킹 시스템이 공고히 작동했으며, 이는 사진이나 동영상 뉴스도 마찬가지였다. 설령 잘못된 정보가 취득되었다 하더라도 사실관계에 대한 엄격한 검증의 단계를 거치면서 최종적으로는 수정·보완된 양질의 뉴스가 대중들에게 전달될 수 있었다. 아울러, 이러한 언론인들의 사실관계 확인 노력이 높은 수준으로 신뢰받은 덕택에 뉴스가 재전달되더라도 그 출처에 대한 정보가 유지되며, 상당 부분 신뢰성을 지켜나갈 수 있었다.

반면, 정보의 흐름이 거미줄처럼 얽혀있는 SNS 플랫폼에서는 소셜미디어가 단순한 소통통로에 그치는 것이 아니라 주체적으로 정보를 생산하고 가공하는 역할까지 함께 하기에 다수에 의한 정보 생산과 더 확산된 다수의 정보 소비가 동시에 이루어진다. 이 과정에서 팩트체킹의 책임과 지위를 부여받았던 정보생산자의 개념은 희박해지고, 결국 전달자와 재전달자만 남아 콘텐츠가 확산되는 경향이 두드러진다. 이러한 플랫폼 특성은 결국 딥페이크 이미지가 무차별적으로 전파되는데 유리한 토대로 작용한다.

덧붙여, 이러한 확산이 편향된 관계를 바탕으로 이뤄지는 것도 위험 요소다. 영상의 진위나 신뢰도가 이미지 자체의 속성에 의해 판단되지 않고 이를 전달하는 채널(전달자)의 인기(팔로워나 조회수)나 호감도에 의해 결정되기가 쉽기 때문이다.[17] 이는 설령 딥페이크로 의심되는 영상이더라도 채널의 신뢰도에 가려 검증과정이 생략될 가능성이 농후함을 의미한다.

SNS 미디어의 기술적인 특성도 딥페이크 확산의 위험성을 가중시킨다. 이용자들이 시·공간의 제약 없이 소셜미디어에 접근할 수 있게 하는 '이동성'[18]과 정보의 수·발신이 즉각적으로 이루어지며 단기간에 여론 형성을 가능하게 하는 '즉시성'은 양질의 정보에만 해당되는 것이 아니라 거짓 정보에도 마찬가지로 적용되는 양날의 검이라 할 수 있다.

마지막으로, 최근 들어 SNS 상에 전달되는 모든 콘텐츠의 길이와 분량이 짧아지는 것도 유념해볼 대목이다. 독해에 시간이 소요되는 텍스트 대신 한 장의 사진이나 짧은 영상이 더 활발하게 공유되고 있으며,[19] 줄어든 영상의 길이는 앞뒤 내용의 분간을 어렵게 해 딥페이크 영상이 검증되는데 필요한 최소한의 정보마저 감추는 효과를 불러오고 있다.

라. 딥페이크와 디지털 민주주의, 선거 리스크

이처럼 건강하고 진실된 여론 형성을 가로막는 딥페이크는 더 나아가 디지털 민주주의의 실현에도 악재로 작용한다. 지난해 벨기에의 한 정당[20]은 정부의 기후변화협약

17) 황유선 (2013). 뉴스 소스로서의 SNS 역할과 그 문제점. 〈언론중재〉, 제128호, p.90~107.

18) Pick, Y. C. (2010). Mobile strategies in political communication. 〈MA thesis〉. American University Washington DC.

19) 대표적인 플랫폼이 15초 내의 짧은 영상을 올려 공유하는 '틱톡(TikTok)'인데, 틱톡에 접속하면 맨 먼저 보게 되는 문구가 '숏·확·행, 짧아서 확실한 행복' 이다.

20) 중도좌파성향의 다른사회당(Socialistische Partij Anders).

[그림 9] 딥페이크로 제작한 트럼프 미 대통령 기후협약 탈퇴 선언 영상

이행을 촉구하는 충격요법 차원에서 트럼프 미국 대통령이 '나는 파리기후협약을 탈퇴한다. 벨기에도 그러리라 믿는다'고 말하는 내용의 딥페이크 영상을 트위터에 올렸다. 문제는 이 영상이 딥페이크라는 사실 표기를 소홀히 해[21] 많은 시민이 트럼프를 비난하고, 정부 차원의 행동을 촉구하는 댓글을 다는 등 해프닝을 넘어선 혼돈이 발생했다는 점이다. 이 사례에서 단편적으로 드러나듯 딥페이크 영상은 정치적 선동의 측면에서 매우 매혹적인 도구이며 점차 완성형으로 진화하고 있는 디지털 민주주의의 실현에 큰 걸림돌이다.

21) 의도적이었다는 의혹이 짙은데, 딥페이크임을 밝히는 부분의 자막을 달지 않았다.

덧붙여, 대의제 민주주의 한계를 극복하고 디지털 민주주의를 실현해 나가기 위한 기술적 토대로서 인공지능이나 블록체인, IoT(사물인터넷) 등이 첨단화되어가는 흐름을 봤을 때, 딥페이크는 첨단 의사결정 시스템의 완성을 저해하고 정책 결정의 합리성과 능률을 떨어뜨리는 요소이다. 알고리즘에 의한 민주주의를 의미하는 알고크러시(algocracy)나 디지크러시(digicracy) 시대가 도래하면서 정책의 결정과 입법 과정이 오프라인에서 온라인으로 대체될 가능성도 충분하지만,[22] 이러한 상승 변화는 네트워크를 바탕으로 한 소통과 의견표현이 진실해야 한다는 대전제를 내포하기에 딥페이크는 이 과정에서도 반드시 넘어야 할 장애물이다.

특히, 한정된 기간 안에 상당량의 정보가 유통되고 이를 바탕으로 국민의 선택이 실현되는 선거와 같은 정치 이벤트에서는 더욱 각별한 유의가 필요하다. 오프라인 공간에서의 대중유세나 매스미디어에 집중했던 과거와 달리, 2000년대 들어 모든 정당은 SNS를 이용한 메시지의 전파와 확산을 최우선으로 고려하고 있으며, 인공지능과 빅데이터에 기반한 디지털 개리맨더링(digital gerrymandering)[23]이 주요 이슈로 떠올랐기 때문이다.

이 때문에 오는 11월 대선을 치르는 미국을 비롯해 많은 국가가 선거 과정에서 딥페이크에 의한 가짜뉴스가 확산하지 않도록 규제하는 각종 법안[24]이나 장치를 준비하고 있고, 한국에서도 선거 기간 딥페이크 영상이 등장해 단시간에 퍼질 수 있다는 경고가 국회 입법조사처를 통해 나오기도 했다.

04 딥페이크 논란 해결을 위한 노력

이처럼 딥페이크 영상의 사회적 해악이 다양한 측면에서 부각되면서 이를 막기 위한 여러 방법이 시도되고 있는데 이는 크게, 1. 기술적 차원에서의 접근, 2. 저널리즘 차원의 노력, 3. 법·제도 차원에서의 규제로 구분해볼 수 있다.

22) Bryan Ford (2020). Technologizing Democracy or Democratizing Technology? : A Layered-Architecture Perspective on Potentials and Challenges. 〈Digital Technology and Democratic Theory〉. University of Chicago Press. p.5.

23) 온라인상에서 정치 성향이 비슷한 사람들끼리 진영논리에 따른 정보를 주고받는 과정에서 확증 편향된 정보에 선택적으로 노출되는 경향이 짙어지고 있으며, 이를 바탕으로 지리적 차원이 아닌 정치 성향에 따른 구역화가 이뤄진다는 의미. Jonathan Zittrain (2014. 6. 20). Engineering an Election: Digital Gerrymandering Poses a threat to democracy. 〈Harvard Law Review Forum〉. URL: https://harvardlawreview.org/2014/06/engineering-an-election/

24) 대표적인 것이 2019년 9월 상원에서 통과된 '딥페이크 보고법(S.2065-Deepfake Report Act of 2019)'이다.

가. 딥페이크 탐지 기술 개발

딥페이크 영상은 일단 유통되기 시작하면 순식간에 무차별적으로 확산되기에 탐지 툴 개발을 통해 사전에 차단하는 것이 매우 중요하다. 다른 차원의 노력은 모두 확산 이후의 대처에 해당하는 것으로 이미 가혹한 피해를 남긴 뒤일 가능성이 높기 때문이다. 수많은 데이터가 오가는 온라인 공간에서 딥페이크 영상물을 탐지해 내기란 쉽지 않은데, 미 국방부 고등연구계획국(DARPA)의 미디어포렌식팀을 비롯해 여러 IT기업과 연구소들에서 딥페이크 영상을 찾아낼 기술을 고민하고 있다.

예를 들어, 딥페이크 영상이 주로 눈을 뜬 사진을 바탕으로 합성되기에 눈 깜빡임이 부자연스럽다는 점에 주목한 탐지방식이나 입술의 움직임과 영상의 목소리가 일치하는지를 검증하는 방식, 원래 얼굴과 바뀐 얼굴의 이미지 해상도 차이를 측정하는 방식 등 인공지능을 활용한 다양한 기술적 시도들이 행해지고 있다.[25)]

아울러, 마이크로소프트는 자사의 클라우드 플랫폼을 통해 조작 영상의 가능성

25) 앞의 박준 · 조영호 (2019).

을 수치화해 신뢰점수지표로 보여주는 딥페이크 탐지프로그램 '마이크로소프트 비디오 어센티케이터(Microsoft Video Authenticator)'를 발표해[26] 얼마 후 진행될 미 대선 과정에서 딥페이크 영상의 폐해를 줄이겠다고 공언했고, 구글 역시 리버스엔지니어링(reverse engineering)을 활용한 '클레임리뷰(ClaimReview)' 툴을 통해 이미지 펙트체크 서비스를 제공하겠다고 밝힌 바 있다. 이외에도 포토샵과 프리미어 등의 이미징 툴 개발사인 어도비는 뉴욕타임스, 트위터와 공동으로 사진·동영상의 원 출처를 확인하게 해주는 디지털 워터마크를 시행 중이며 움직이는 이미지(gif) 사이트인 지피캣은 각각 앙고라(Angora)와 마루(Maru) 프로젝트를 통해 조작된 이미지를 탐지하는 기술을 연구하고 있다.

나. 저널리즘 영역의 노력

딥페이크 영상물에 대한 위기의식이 고조되며, 가짜 이미지 뉴스를 퇴출시키려는 언론계 내의 노력도 활발하다. 대표적인 움직임은 유럽의 유수 언론사가 참여하고 있는 인비드(Invid) 프로젝트[27]다. 프랑스의 통신사 AFP와 독일의 도이체벨레(Deutsche Welle)를 포함해 여러 대학과 IT 보안업체들이 참여한 이 프로젝트는 온라인 공간에서 유통되는 영상물에 딥페이크로 제작된 영상이 있는지 탐지하는 프로그램을 개발, 웹브라우저 플러그인과 모바일 앱으로 제공하고 있다.

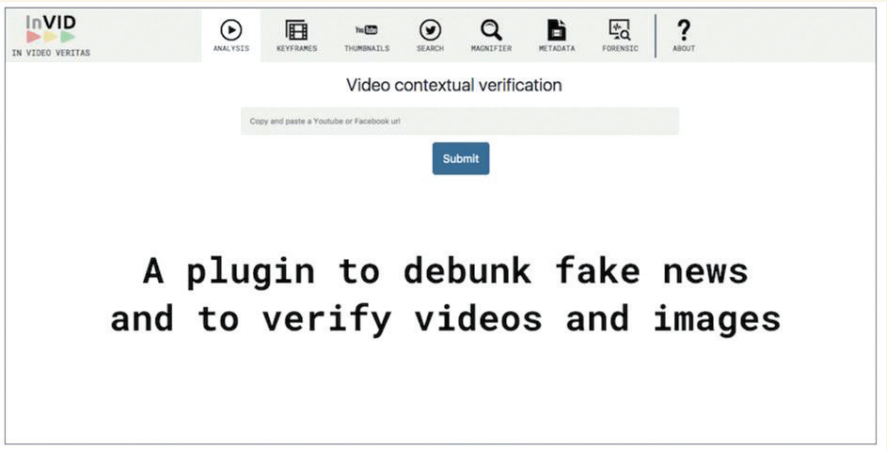
프랑스의 공영방송 프랑스24는 5천여 명의 시청자와 함께 '옵저버(The Observer)' 프로그램을 운용하는데, 팩트체크에 관심 있는 사람들이 온라인상에서 유통되는 영상물을 직접 검증하는 일종의 시민 참여형 팩트체크 프로그램이다.[28] 우리나라에서도 30여 곳의 언론사가 참여한 가운데, 서울대 언론정보연구소가 운영하는 'SNU팩트체크센터'가 논란이 되는 가짜뉴스들을 검증하며 활발한 활동을 이어가고 있다.

이 밖에도 전 세계 언론인이 함께 모여 가짜뉴스를 퇴출시키기 위한 방안을 모색하는 모임도 해마다 열리고 있는데, 55개국 251명의 언론인과 팩트체커가 참여한 가운데 지난해 남아프리카에서 열린 '6차 세계팩트체킹서밋'에서는, 딥페이크 영상의 문제점이 주요 화두로 등장해 이를 탐지하기 위한 참신한 해법들이 논의되었다.

---

26) 참고 URL: https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/
27) 참고 URL: https://www.invid-project.eu
28) 옵저버 프로그램의 모토는 'Film, Verify, Share(촬영과 검증, 그리고 공유)'이다.

[그림 10] 인비드(InVID)가 지원하는 가짜이미지 검증 플러그인(Plug-In)



다. 법/제도적 차원의 규제

기술, 언론 영역의 노력과 별개로, 딥페이크 영상물 제작·유포행위를 처벌하는 규제 법안의 마련도 시급하다. 우리나라에서 딥페이크 영상은 음란물의 차원에서 초기 사회문제화되었기에 규제법안 역시 이 부분에 초점이 맞춰져 있지만 앞서 살펴본 바대로 저널리즘과 정치 참여의 영역에까지 심각한 위협이 가해지고 있는 실정에서 좀 더 포괄적인 대응책이 필요한 상황이다.

올 초 사회를 떠들썩하게 했던 '텔레그램 N번방 사건' 이후 국회는 불법 촬영물 유통 방지와 관련한 입법 조치를 마련했는데, 여기에는 통신사업자의 불법 촬영물 유통방지 의무를 강화하는 내용은 물론, 카카오, 네이버, 구글 등 부가통신사업자에게 불법 촬영물 등의 삭제·접속차단 등 유통방지조치 의무도 부과하고 있다. 입법 보완 차원에서 한 가지 방법은 이 법안의 '불법 촬영물' 범주에 딥페이크로 조작된 영상물을 포함시켜 법 조항 적용의 폭을 넓히는 방안을 생각해볼 수 있다.

하지만 아동 성착취물처럼 유해성이 선명한 콘텐츠가 아닌 경우 처벌조항의 적용이 모호해질 수밖에 없으며, 풍자와 유희의 목적으로 만들어지거나 밈(Meme)문화 차원의 딥페이크 영상[29]은 그 자체로 순기능을 포함하고 있기 때문에 법의 테두리에서 논의될 근거가 없다. 때문에 보다 확장된 개념 정리가 필요한데, 이와 관련해 최근 외국에서 채택되고 있는 몇몇 법안들은 이 논의의 실마리를 찾을 수 있게 해준다.

29) Tormod dag Fikse (2018). Imaging Deceptive Deepfakes An ethnographic exploration of fake videos. 〈MA thesis〉. University of Oslo

미국의 경우, 지난 해 딥페이크 책임법(Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019 : H.R. 3230)[30]이 발의되었는데, 이 법안은 딥페이크의 범위와 적용을 상세히 규정함과 동시에 선거 과정에서 발생하는 딥페이크 범죄에 대처하기 위해 국토안보부 산하에 별도의 TF팀을 구성할 것을 명시하고 있다. 아울러 캘리포니아에서도 지난해 10월 딥페이크 영상을 금지하는 법안이 통과되었는데, 개빈 뉴섬 주지사는 정치인의 가짜 발언이나 행동이 담긴 딥페이크 영상(오디오)을 제작하거나 배포할 경우 범죄로 간주하는 'AB730' 법안에 서명했다. 이 법안은 선거 전 60일부터 모든 후보자에게 적용되며 단, 뉴스매체의 풍자나 패러디, 영상 안에 페이크임을 밝힌 경우는 예외로 인정된다.[31]

유럽 역시 딥페이크 여부를 세세히 구분해 대처하고 있다. 유럽연합(EU)은 2018년 딥페이크를 포함한 허위정보에 대처하기 위해 '허위정보행동규범(Communication-Tackling online disinformation : a European Approach)'을 발표했는데, 이 법안은 허위정보식별을 위한 출처 표기를 강화하고 신뢰도 여부를 파악하는 조치를 담고 있으며 인공지능 플랫폼에 대한 제3자 검증기능도 강화하고 있다. 이와 별도로, 프랑스는 '정보조작대처법', 독일은 '네트워크 집행법' 등을 통해 딥페이크 조작을 처벌할 근거를 마련했다.

30) 참고 URL: https://www.congress.gov/bill/116th-congress/house-bill/3230/text
31) 참고 URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

05 결론

최근 저널리즘 영역의 떠오르는 화두 중 하나는 인공지능 기자의 뉴스 생산이다. AP 통신이 시행 중인 오토메이티드 인사이트(Automated Insight)나 포브스지가 도입한 내러티브 사이언스(Narrative Science)의 사례에서 알 수 있듯, 인공지능 기자가 생산한 콘텐츠는 인간이 작성한 기사와 구분되지 않을 정도로 완성형이며 스포츠 중계나 기상정보 등 로봇 기자가 작성한 뉴스가 이미 대중들에게 전달되고 있다. 저널리즘 역시 인공지능 기술의 한복판에 들어와 있는 것이다.

한편, 이러한 자동화 양상은 팩트체크와 공정성 확보의 노력이 인간의 손에서 벗어날 수도 있음을 의미한다. 딥페이크의 고도화가 위협적인 것도, 미디어의 생산과 소비 패러다임이 급격히 전환되는 것과 맞물려 있다. 기술의 발전이 늘 제도를 앞질러 가듯, 각종 규제법안과 장치의 마련에도 불구하고 딥페이크 이미지는 점점 더 그럴듯하게 꾸며져 대중들에게 전달될 수 있으며, 이제는 이러한 위험성의 상존을 늘 감안해야 하는 새로운 저널리즘의 시대에 들어서고 있는 것이다.

상술한 모든 노력에 우선해, 수용자 중심의 이미지 리터러시를 키우는 것이 가장 중요하다. 이미지 리터러시란, 매체 해석능력을 의미하는 미디어 리터러시 개념을 조금 더 세분화한 것으로 조작된 사진이나 영상 등 이미지를 비판적으로 검증하고 진위를 확인하는 능력을 의미한다. '보는 것이 믿는 것'이던 시대는 저물었고, SNS 시대 정보 소비자는 동시에 정보 생산자이기에 모든 단계에서 팩트체킹에 대한 책임감을 가질 필요가 있다. 딥페이크로 조작된 이미지가 노리는 것이 인물의 대체를 통한 정체성의 도둑질이라면, 그러한 장물이 거래되지 않는 시장 환경을 만드는 것은 결국 언론 소비자의 역할이다. 그리하여 궁극적으로는 이미지 자체의 '사실성'을 넘어, 이미지 '공정성'까지도 확보하는 포괄적인 저널리즘 토대가 구축되도록 노력해야 한다.

전 세계가 코로나 바이러스로 몸살을 앓고 있는 지금, 많은 미래학자들은 코로나 이후의 세계가 기존과는 다른 기준이 통용되는, 이른바 '뉴노멀'의 질서가 자리 잡을 것이라 예견한다. 비유하자면, 저널리즘 역시 이미지에 대한 사실확인과 교차검증을 습관화하는 이미지 팩트체크의 시대 '뉴노멀'을 준비해야 한다. 기술–제도–소비자 3주체의 노력이 결실을 맺는 것도 이러한 바탕 하에 비로소 풍성해질 수 있다.