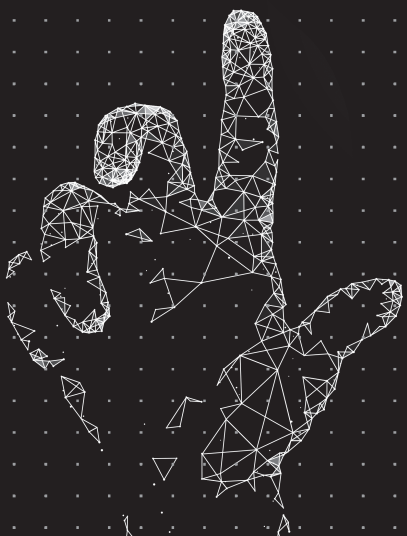
1

인류를 위협하는 미지의 재앙 X이벤트



1

첨단기술의 역습

KAIST Future Strategy 2022

+ 유권자를 조종하는 디지털 프로파간다

+ AI 알고리즘의 오작동

+ 유전자 가위 기술에 의한 차별적 미래 사회

+ 인간 뇌와 AI 결합의 가능성과 위험

+ 진짜 같은 가짜, 딥페이크의 위협

유권자를 조종하는 디지털 프로파간다

최근 인공지능AI 알고리즘의 스토리텔링 기술이나 딥페이크deepfake 기술이 소셜미디어SNS를 통해 광범위하게 활용되면서 이를 이용한 허위·조작 정보disinformation와 가짜 뉴스fake news가 큰 문제가 되고 있다. 이를 만들어내는 주체들이 다른 국가의 민감한 정치 상황이나 선거 여론전에까지 개입해 국제 정치적 논란을 낳기도 한다. 미국의 2016년 대통령 선거를 비롯한 각종 서구권 선거에서 퍼졌던 허위·조작 정보의 출처가 러시아 정부와 연계된 정보전information warfare으로 드러난 것이 그 예다. 이러한 '디지털 프로파간다digital propaganda' 혹은 '사이버 심리전'은 정상적인 여론 형성을 방해하면서 민주주의의 가치와 제도를 심각하게 훼손하고 있다. 만약 대통령 선거 유세 기간에 다음과 같은 일이 발생한다면 어떻게 될까?

가상 시나리오: 대선을 좌우한 정치 유튜버의 죽음, 진실인가, 거짓인가?

범죄심리학을 전공한 A, 소설가인 B, 정치평론가인 C는 모두 러시아와 가까운 동유럽 국가에서 인기를 누리는 유명 유튜버다. 다가오는 자국 대통령 선거에서 친러 성향의 후보를 지지한다는 공통점이 있다. 이들은 대선 캠페인이 시작되자 일제히 선거 이슈들을 다루기 시작하면서, 특히 이들의 성향과 반대인 친EU 성향의 후보를 집중적으로 공격한다. 이들이 소개한 악의적 가짜 뉴스는 소셜미디어를 통해 빠르게 퍼져나간다.

선거일 사흘 전부터 이들은 친EU 후보 캠프와 미국으로부터 자신들이 살해 위협을 받고 있다고 호소한다. 역술가 유튜버인 D가 이들의 사망을 예언하면서 세간의 이목은 더욱 집중된다. 그리고 놀랍게도, 선거 하루 전 A와 B가 사망하고 C는 실종된다. 이에 분노한 친러 후보 지지자들이 친EU 후보의 사퇴를 요구하고 EU 가입을 반대하는 대규모 시위를 벌인다. 결국, 대통령 선거는 친러 후보의 승리로 끝난다.

그런데 경찰의 조사 결과는 매우 충격적이다. A, B, C의 모습이 담긴 영상이 모두 딥페이크 기술로 만들어진 조작 영상이며, 세 명 모두 실존하지 않는 가상 인플루언서virtual influencer라고 발표한 것이다. 또 이들의 유튜브 채널은 러시아의 한 지역에서 운영되었지만, 러시아의 비협조로 수사가 어렵다고 한다. 그러나 D는 이러한 경찰 조사 결과를 믿을 수 없다며 경찰이 세 명의 시신을 모두 숨기고 있는 것이라고 자신의 유튜브 채널에서 주장하는 중이다. 이 중 무엇이 진실일까?

AI 알고리즘의 최첨단 프로파간다 전술

인류 역사에서 프로파간다는 가장 빈번하게 수행되어온 대중 설득 전략이다. 프로파간다란 "목표 청중의 감정, 태도, 의견, 행동에 영향을 끼치려는 체계적 형태의 의도적 설득 행위로서, 이념적·정치적·상업적 목적을 위해 미디어를 통해 통제된 방식으로 메시지를 일방향으로 전달하는 것"[1]을 말한다. 프로파간다라는 용어가 부정적인 이미지를 갖게 된 것은 중세 시대 가톨릭교회가 프로테스탄트의 주장을 프로파간다로 묘사하면서부터였다. 이후 제2차 세계대전 당시 나치 독일의 악명 높은 활동으로 인해 프로파간다는 '거짓말', '조작', '왜곡', '정신 조종', '심리전', '기만', '세뇌' 등과 유사하게 사용되면서 부정적인 의미가 고착되었다.

오늘날 사회적으로 논란이 되는, 주로 인터넷과 소셜미디어를 통해 전개되는 프로파간다 활동이 '컴퓨터 프로파간다' 혹은 '디지털 프로파간다'로 불리는 이유는 온라인 공간에서의 정보 생산과 확산에 AI 알고리즘과 같은 첨단 디지털 기술이 본격적으로 사용되고 있기 때문이다. 주로 소셜미디어의 가짜 계정을 통제하는 대화형 AI 알고리즘 프로그램인 소셜봇social bots은 정치적 목적을 위해 동원되기도 하므로 '정치봇political bots'이라고도 불린다.

이러한 AI 봇들은 다양한 프로파간다 전술을 펼친다. 예를 들면 특정 키워드나 저명한 정치인의 트윗과 같은 촉발 변수trigger를 사용해 특정 정보를 대규모로 유포하기도 하고, '팔로워 봇follower bots' 같은 경우 가상 인물을 만들어 팔로워 수를 늘리고 대규모의 '좋아요'를 생성하기도 한다. 또 '길 차단 봇roadblock bots'은 온라인 공간에서 사람들의 주의를

다른 이슈로 분산시켜 관심을 돌려놓거나 영향력 있는 사용자의 게시물에 스팸성 해시태그를 다량 게시해 원래의 글을 밀어낸다. 이 밖에도 특정 인물에게 협박 메시지를 보내거나 해킹으로 개인 정보를 유출하기도 하고, 선거 감시 웹사이트나 모바일 앱의 작동을 방해하는 등 극단적 방식으로 프로파간다 활동을 수행할 수 있다.

프로파간다의 설득 전략은 다양한 심리학 이론에 근거한다. 예를 들어 '진실 착각 효과illusory truth effect'가 있다. 사람들이 처음 접했을 때 신뢰하지 않았던 정보원의 정보라도 시간이 지나면 불확실한 정보원에 대한 기억은 잊고 그 정보를 기정사실로 인식하는 경향이다. 다시 말해 허무맹랑한 메시지에 반복적이고 빈번하게 노출되면, 사람들은 그 메시지를 점차 신뢰하게 되며, 처음에 접한 정보를 나중에 접하는 정보보다 더 믿고 따르게 된다는 것이다.

디지털 프로파간다의 위험성

최근 인지신경과학 연구에 따르면, 인간이 지닌 놀라운 능력 중 하나는 세상의 온갖 다양한 일 가운데 '예상 밖의 정보나 사건을 신속하게 탐지하는 능력'이라고 한다. 인간의 인지력은 항상 '새롭고 참신한 것novelty'에 주목하고 집중하는 인지 방식을 길러왔으며, '새로운 자극novel stimuli'을 경험할 때마다 뇌에서 도파민이 증가해 일종의 보상을 제공하고, 결과적으로 학습과 기억을 돕는다는 것이다. 이러한 이유로 인간의 뇌는 감정적이고 도발적인 정보를 더 오랫동안 강하게 기억하고, 그런 측면에서 프로파간다 목적을 갖는 가짜 뉴스 정보는 '거짓 기

억false memories' 현상을 쉽게 유발할 수 있다. 거짓 기억 현상이란 실제로 일어나지 않은 일을 일어났던 일로 기억하고 믿어버리는 상태를 말한다.

한 실험에 따르면, 가짜 뉴스 정보가 개인의 믿음이나 신념과 관련될 때 사람들은 그러한 거짓 기억 경향을 더 쉽게 보였다고 한다. 가짜 뉴스의 힘은 바로 사람들의 기억을 사로잡고 학습을 촉진할 만큼 이목을 끄는 내용에 있고, 이때 '도덕적 감정moral emotion'을 강하게 자극하는 정보일수록 더 크게 주목받으며 소셜미디어 공간에서 빠르게 확산하는 경향을 보인다. 격한 반응을 일으킬 만한 내용의 가짜 뉴스는 그만큼 사람들이 세상을 인식하고 정치적 결정을 내리는 데에 강한 영향력을 발휘한다.

가짜 뉴스가 퍼질 때 그러한 정보에 가장 먼저 반응하는 그룹은 뉴스 미디어와 관련 분야의 전문가들이다. 하지만 가짜 뉴스가 퍼지는 속도와 범위만큼 팩트 체크가 이뤄진 정확한 정보가 전달될지는 미지수다. 무엇보다 팩트 체크가 이뤄진 온전한 정보가 가짜 뉴스만큼 사람들의 주목을 받으며 흥미를 유발할 수 있을지는 더 불확실하다.

사실 확인 과정에서 더 문제가 되는 것은 허위·조작 정보를 퍼뜨린 사람이나 세력이 팩트 체커인 양 가장한 채 또 다른 허위·조작 정보를 만들어낼 수 있다는 것이다. 허위·조작 정보를 생산하고 소셜미디어에서 퍼뜨리는 행위자가 갖는 동기는 대개 정치적이거나 경제적이다. 그는 사용자의 주목을 받을 수 있는 콘텐츠를 끊임없이 제작·공급하려고 할 것이고, 그가 만드는 허위·조작 정보의 자극성과 선정성 또한 더욱 높아질 수밖에 없다.

가상의 전장이 된 인터넷

모든 허위 정보misinformation가 허위·조작 정보는 아니며 고의 없이 잘못된 정보false information를 우발적으로 공유하는 것도 허위·조작 정보가 아니다. 가짜 뉴스와 같은 허위·조작 정보는 누군가를 오도하기 위한 목적으로 잘못된 정보를 유포하는 활동이다. 특히 딥페이크 영상은 AI 알고리즘을 이용해 동영상 원본의 사람을 다른 사람의 모습으로 편집한 뒤 편집된 인물이 실제 존재하는 것처럼 조작하는 기술이 동원되는, 진위 식별이 가장 어려운 형태의 허위·조작 정보다.•

한편, 정치적으로 악의적인 목적을 위해 만들어진 허위·조작 정보가 소셜미디어를 통해 대규모로 유포되면서 심각한 국제 정치적 논란을 일으킨 것은 2016년 미국 대통령 선거 때부터다. 이때부터 미국과 유럽의 거의 모든 선거 기간에 러시아 정부와 연계된 것으로 밝혀진 허위·조작 정보가 소셜미디어를 기반으로 퍼져나갔다. 러시아의 이러한 공격적인 사이버 심리전은 서구권의 여론 분열을 통해 선거 결과에 직접 영향을 끼치고 민주주의 제도를 훼손하려는 시도이자 주권 침해 행위로 비난받고 있다. 디지털 프로파간다로 일컫는 이러한 활동이 위협적인 이유는 서구권 사회의 여론 왜곡과 사회 분열에 그치지 않고, 공격 대상 국가의 의사결정 과정을 마비시키고 정부의 정치적 정당성과 민주주의

• 물론 딥페이크가 악의적 목적으로만 사용되지는 않는다. 예컨대 미켈라 소사Miquela Sousa는 2016년부터 릴 미켈라Lil Miquela라는 닉네임으로 소셜미디어에서 활동하고 있는 버추얼 인플루언서다. 자신이 딥페이크 기술로 만들어진 가상의 인물이라고 밝힌 미켈라는 수많은 인스타그램 팔로워를 거느리고 있으며, 패션 모델로 인기를 누리고 있다. 이마Imma, 리암 니쿠로Liam Nikuro, 실비아Sylvia 등도 유명한 버추얼 인플루언서다.

제도까지 무력화하는 것을 목적으로 삼기 때문이다. 더군다나 이러한 사이버 심리전은 전시와 평시를 가리지 않으며 평시에는 더욱 은밀하게 수행될 수 있어 허위·조작 정보를 이용한 상시적 사회 분열 시도가 가능하다. 이에 따라 미국과 유럽은 이를 '사이버 테러'로 간주하고 군사 안보 차원에서 엄중히 대응하고 있다.

앞서 언급한 가상 시나리오는 극단적인 상황을 그려본 것이지만, AI 알고리즘의 정보 통신 기술을 이용하면 결코 불가능한 일이 아니다. 특히 이념이나 이슈를 놓고 분열된 사회일수록, 진영 대결로 인한 분노와 증오로 가득한 사회일수록 자극적인 가짜 뉴스에 취약해진다. 요컨대 현대의 디지털 첨단 사회에서 인터넷이나 소셜미디어 공간은 AI를 활용한 정보 통신 기술을 통해 언제든지 치열한 가상의 전장으로 전환될 수 있게 된 셈이다.

온라인 공론장이 사이버 심리전에 무너지는 이유

2016년 미국 대선을 기점으로 서구권은 주요 선거 때마다 가짜 뉴스를 이용한 러시아의 사이버 심리전 공격에 휘둘렸다. 2016년 영국 브렉시트 국민 투표, 2017년 프랑스 대선, 독일 총선, 스페인 카탈루냐 독립 투표, 2018년 이탈리아 총선, 2019년 유럽 의회 선거, 2020년 미국 중간선거와 대선 등등. 서구 민주주의의 온라인 공론장은 왜 그리 무력하게 당했던 것일까?

가장 큰 이유는 앞서 짚어본 것처럼 사이버 심리전에 이용된 스토리텔링 설득 전략이 '인지적 해킹' 혹은 '정신적 해킹'으로 일컬을 정도로

상당히 정교하게 고안되었기 때문이다. 또 대규모 봇 부대가 자극적인 가짜 뉴스를 특정 기간에 대규모로 빠르게 유포했기 때문이다. 정치 논쟁이 활성화되고 사람들이 쉽게 양극화되는 선거 유세 기간에는 정치적으로 민감한 허위·조작 정보가 수많은 유권자에게 지대한 영향을 끼칠 수 있다. 2016년 미국 대선 결과에 대한 〈포브스〉 지의 보도에 따르면, 선거일이 임박할수록 유권자들은 주류 뉴스보다 가짜 뉴스에 더 주목했을 뿐만 아니라 정확한 뉴스와 가짜 뉴스를 비슷한 수준으로 믿었다. 선거철 가짜 뉴스가 유권자들의 투표에 결정적 영향을 끼칠 가능성이 매우 크다는 점을 시사한다.

러시아와 같은 권위주의 정권이 수행하는 사이버 심리전의 전쟁 방식이 민주주의 사회에 갖는 강점은 비대칭전, 그리고 비정규전이라는 점이다. 개방된 민주주의 사회의 온라인 공론장에 대한 교란 전술은 상대적으로 적은 비용으로 선제 공격의 우위를 점할 수 있고 원하는 시기에 언제든지 게릴라전처럼 수행할 수 있는, 매우 가성비 높은 공격 형태다. 최근 중국도 중국어권을 타깃으로 한 디지털 프로파간다 활동을 본격화하고 있다. 2019년 홍콩의 범죄인 인도 법안에 대한 시민들의 반대 시위 및 2020년 타이완 총통 선거와 입법위원 선거에서 중국은 소셜미디어 공간을 통해 중국어로 된 가짜 뉴스를 대규모로 퍼뜨린 바 있다.

하지만 사이버 심리전의 고도로 정교한 전술을 이해한다고 하더라도 서구의 온라인 공론장이 그렇게 취약하다는 사실은 여전히 해소되지 않는 의문을 남긴다. 냉전기에 그들이 공산 진영의 프로파간다 전술에 말려들지 않고 자유주의 이데올로기를 지켜냈던 것을 떠올리면 더욱 그렇다. 그 취약성의 이유는 러시아가 취한 사이버 심리전의 교란 전술이 서구 민주주의 사회의 약점을 파고들었기 때문이다. 계속되는 세

계적인 경제 침체, 각국의 경제적 양극화, 포퓰리즘과 극우 민족주의의 득세는 민주주의 사회의 내부 갈등을 증폭시켰고, 바로 이 지점에 사이버 심리전의 공격이 집중되었다. 서구권 사회에서 나타나고 있는 사회적 균열과 파괴되어가는 사회적 연대 및 약화된 민주주의가 온라인 공론장을 위기로 몰아간 원인이다.

사이버 심리전 공격에 대한 미국과 유럽의 대응

미국과 유럽은 AI 알고리즘 기술을 이용한 러시아와 이란 등의 사이버 심리전 공격을 자신들의 주권과 민주주의 제도에 대한 중대한 도전으로 간주하고 군사·안보 차원의 대응 체계를 마련하고 있다. 일례로 미국은 국무부의 공공 외교 및 공보 담당 차관 산하에 글로벌 인게이지먼트 센터Global Engagement Center를 두고 미국과 동맹의 안보에 영향을 끼치려는 타국 정부 및 비정부 행위자들의 프로파간다 활동과 허위·조작 정보 공작에 대한 대응을 전담하게 하고 있다. 또 국방부 산하 고등연구계획국DARPA에서는 딥페이크 영상을 탐지하는 미디어 포렌식 연구를 진행 중이며, 세계적인 미국의 IT 업체들과 협업하고 있다.

유럽은 2018년부터 디지털 허위·조작 정보 대응을 위한 '허위·조작 정보 대응 행동 규약Code of Practice on Disinformation'을 발표하고 민간의 자율 규제를 권고해왔다. 그러나 최근 강력한 규제를 하는 방향으로 정책이 변화하고 있다. 북대서양조약기구(나토NATO)의 경우 사이버 심리전 대응을 위해 2014년 라트비아에 나토 전략 커뮤니케이션 센터를 설립했다. 이 기관은 다국적군, 민간, 학계 전문가로 구성되었으며 나토와

우방국 간의 전략 커뮤니케이션 능력을 강화하고 역내 사이버 공격과 정보전 정책을 지원한다. 나토는 허위·조작 정보 유포 상황을 가정한 다양한 시뮬레이션 훈련도 전개하고 있다.

그동안 많은 국가가 사이버 공간에서 발생하는 다양한 문제를 주로 사이버 공격과 관련된 기술적·법적 이슈의 측면에서만 다뤄왔다. 하지만 디지털 프로파간다는 사회 전체 네트워크와 시스템에 대한 도전의 성격을 띠고 있는 만큼 포괄적 차원에서 대응해야 한다. 마찬가지로 국내 미디어에서 유통되는 오정보나 정치적으로 편향된 정보를 해외 언론 등이 무분별하게 번역해 인용하는 경우, 한국과 관련된 가짜 뉴스나 잘못된 정보가 해외 소셜미디어 공간에서 퍼지기도 한다. 이는 단순히 기술적인 측면에서만 다룰 문제가 아니라 정보 커뮤니케이션의 윤리와 규범 등도 포함해 종합적으로 다뤄야 하는 부분이다.

AI 알고리즘의 오작동

인간과 인공지능AI 간의 세기의 대결이라 불린 2016년 이세돌 대 알파고의 바둑 대국은 역사적 사건이었다. AI 바둑 프로그램 알파고AlphaGo가 압승을 거둔 후 AI에 관한 관심은 더욱 높아졌고, 실제로 알파고의 승리는 더딘 발전을 보이던 AI 기술의 획기적 변화를 상징했다. 이제 AI 혹은 지능형 알고리즘은 우리 삶의 곳곳에 스며들어 여러 분야에서 '스마트'한 경험을 제공하고 있다. 그런데 만약 AI 알고리즘이 손상되거나 오작동을 일으키면 어떻게 될까? 악의적인 외부 공격으로 파괴되면 또 어떻게 될까? AI 바둑 프로그램의 오류 정도가 아니라 자율주행차의 AI 알고리즘이 오작동을 일으키고, 국가 핵심 네트워크의 AI 알고리즘이 파괴된다면? 그 피해는 누구도 쉽게 가늠하기 어려울 것이다. 이런 피해는 그 가능성을 아무리 최소화한다고 해도, 일단 발생하면 치명적 상황으로 치달아 'X이벤트'가 되어버릴 수 있다.

극단적 사건을 상상하라

원전에는 '설계 기준 사고design basis accident'라는 개념이 있다. 설계할 때 참조하는 가상의 사고를 뜻한다. 발생 가능성이 희박하더라도 공중의 안전을 위해 처음부터 사고를 염두에 두고 설계를 한다는 의미다. 2011년 지진과 쓰나미로 파괴된 일본 후쿠시마 원자로의 경우, 설계자들은 과거 일본이 경험한 최대 지진인 8.3 규모에 대비해 벽 높이를 설치했다. 그런데 후쿠시마 원전 사고 당시 지진의 규모는 9.0이었다. 결국 쓰나미로 인한 해수가 유지 벽을 넘으면서 원자로가 망가지고 폭발이 일어났으며 대규모 방사능이 유출되는 참사로 이어졌다. 만약 23만 명의 사망자를 낸 2004년 인도네시아 수마트라 대지진(규모 9.1)을 고려해 유지 벽을 더 높이 쌓았더라면 어땠을까? 결과론적 이야기지만, 후쿠시마 원자로 설계자들이 '설계 기준 사고'를 근래 일어난 최악의 지진 규모로 상정해 9.1 이상에 대비하는 유지 벽을 만들었더라면 원전 폭발의 참사를 막을 수 있었을지도 모른다.

그런가 하면 2021년 5월에는 미국 송유관 업체 콜로니얼 파이프라인Colonial Pipeline에 동유럽 해킹 단체 다크사이드DarkSide가 악성코드인 랜섬웨어 공격을 해 송유관이 차단되는 사태가 일어났다. 해킹으로 파이프에 설치된 각종 IoT(사물인터넷) 센서들이 작동을 멈춘 것이다. 콜로니얼 파이프라인은 미국 동부 지역 석유 공급의 45%를 담당하고 있어 그 여파는 엄청났고 비상사태가 선포되기도 했다. 콜로니얼 파이프라인은 500만 달러가 넘는 비트코인을 다크사이드에 입금하고 나서야 시스템을 정상화할 수 있었다. 만약 국가의 중추적 AI 시스템이 이러한 악성코드 공격을 받아 멈추거나 반사회적으로 오작동한다면 그 피해는 감

당하기 어려울 것이다.

이러한 극단적 사건들은 또 일어날 수 있다. 특정 지역을 넘어 전 세계에서 발생할 수도 있다. 이미 전 세계를 덮친 코로나19가 그러했다. 통신 장애, 핵 공격, 소행성 충돌, AI의 오류 등도 감염병 바이러스처럼 우리 사회에 엄청난 재앙이 될 수 있다. 이런 일들을 예측하고 상상해 미리 대비하는 것만이 우리가 할 수 있는 최우선의 방법이다.

일상을 주도할 AI

AI를 가동하는 3대 축은 빅데이터, 컴퓨팅 파워, 알고리즘이다. 컴퓨팅 파워 분야만 해도 성능과 처리 용량, 속도가 빠르게 발전하고 있으며 양자 컴퓨터의 상용화가 가까운 시일 내에 가능할 것으로 보인다. AI를 가동하는 3대 축이 바퀴처럼 맞물려 빠르게 돌아가면서 AI의 지능형 알고리즘이 매일, 매시, 매분 강화되고 있다.

이렇게 발전을 거듭하는 AI 시스템은 가까운 미래에 사람의 개입을 최소화하는 형태로 사회 곳곳에 자리하게 될 것이다. '휴먼 에러human error', '위험한 업무', '복잡성'이라는 세 가지 이유 때문이다.

AI가 사람을 대체하게 되는 첫 번째 이유는 사람의 실수로 인한 사고, 즉 '휴먼 에러'의 발생을 줄이기 위해서다. 자동차 운전 등 휴먼 에러가 발생할 수 있는 영역에서부터 적극적으로 사람에서 AI로의 대체가 일어날 것이다. 두 번째는 '위험한 업무' 때문이다. 원자력 발전소를 비롯해 각종 대형 공장, 전기 케이블을 고치는 일까지 사람이 하기에 위험한 일들을 AI의 지시 아래 로봇, 드론 등이 직접 수행함으로써 인간을 대체

하게 된다. 세 번째, 가장 큰 이유인 '복잡성' 때문이다. 전력, 식수, 식량, 통신, 교통 기관, 의료, 방위, 금융 등 전 영역에서 복잡성이 증가하면서 '복잡성의 과부하' 현상이 일어나고 있다. 2008년 글로벌 금융위기를 불렀던 리먼 브라더스 파산 사태도 인간의 탐욕 때문만이 아니라, 금융 공학자들이 꼬리에 꼬리를 물듯 수많은 금융 파생 상품을 만들어내면서 그 복잡한 연결 고리를 한눈에 파악하지 못했기에 일어난 사건이었다. 이런 복잡성의 과부하는 각 영역에서 잠재적인 문제 발생 가능성을 높여가고 있으며, 이를 통제하고 관리하기 위해 결국 AI가 개입하거나 주도하는 영역이 더 넓어질 것이다.

AI 주도 사회의 위험 요인

이런저런 시행착오를 거듭하면서, 때로는 비효율적인 방식을 통해 서서히 배움을 얻어가는 인간과 달리 AI는 최단 시간 내에 모든 변수를 고려해 시뮬레이션해본 다음 가장 효율적인 경로를 결정한다. 이런 능력은 더 고도화되고 있다. 그러나 AI가 중심이 되는 AI 주도 사회는 예상보다 더 심각한 문제를 가져올 수도 있다.

인간의 배제와 소외

우선 문제해결 과정에서 인간이 소외되는 일이 발생할 수 있다. 로봇 산업을 선도하고 있는 쿠카 로보틱스KUKA Robotics 공장을 가보면 로봇이 로봇을 만든다. 드문드문 사람들이 있지만, 생산 현장에서 예외적인 상황이 벌어졌을 때만 나서서 점검한다. 직원들은 적은 시간 일하고 많

은 휴식 시간을 가지는 근로 형태를 유지한다. 사람이 많이 일하지 않아도 로봇이 대부분 알아서 업무를 수행하기 때문이다. 여기까지는 굉장히 좋은 모델처럼 보인다. 그러나 모든 생산 현장이 원천적으로 사람의 개입 없이 혹은 개입이 배제된 형태로 돌아가게 되면 일자리에서 사람이 소외되는 문제가 발생하기 시작한다.

삼성전자의 반도체 프로세스 맵을 보면 검은 선으로 가득 찬 종이를 볼 수 있다. 그만큼 공정 자체가 복잡하고 정밀하다. 이러한 공정 체계 역시 사람이 컴퓨터의 도움을 받아 발전시킨 것이지만, 앞으로 AI 주도 사회에서는 사람의 개입 없이 AI가 자체적으로 가장 최적의 공정 경로를 만들어낼 것이다. 증가하는 복잡성 때문에 아무리 지능이 뛰어나고 경험이 축적된 사람이라고 해도 AI가 만들어낸 결과물을 이해할 수 없을지도 모른다. 이런 방식으로 대다수 영역에서 인간의 역할과 판단을 배제하는 현상이 심화할 것이다.

AI 알고리즘의 신뢰성 문제

AI는 고도화된 지능형 알고리즘이라고 할 수 있다. 이러한 알고리즘은 '블랙박스'처럼 내용물이 보이지 않는다. 입력과 출력에 해당하는 결과물만 보이고 지능형 알고리즘의 내부에서 어떤 일이 어떤 방식으로 벌어지는지 알 수 없다는 뜻이다.

단순한 예로, 구글의 지도 애플리케이션을 한국에서 열어보면 한국의 동쪽 바다는 '동해'로 표시되어 있다. 반면 일본에 있는 사람이 구글 지도를 열면 한국의 동쪽 바다는 '일본해'로 표시되어 있다. 인도와 파키스탄의 대표적인 분쟁 지역인 카슈미르의 경우도 그렇다. 인도 사람이 보면 인도 땅으로, 파키스탄 사람이 보면 파키스탄 땅으로 보인다. 구글

지도가 민감한 지역 이름을 그 사람이 있는 국가와 장소에 따라 다르게 보여주는 것이다. 실체는 하나인데 사람과 지역에 따라 다른 이름을 보여준다면 객관성과 보편성을 확보할 수 없으므로 AI 알고리즘의 신뢰성에 의문이 들 수밖에 없다.

기계의 도구가 된 인간

2013년 미국 펜실베이니아 지역에 처음 도입된 범죄 예측 시스템인 프레드폴PredPol이라는 프로그램이 있다. 이는 예산 부족으로 순찰 인력이 감소하자 효율적인 순찰 업무를 진행하기 위해 만든 것이다. 이후에 고도화된 지능형 알고리즘이 적용되면서 지금은 실질적으로 범죄를 예측하고 예방하는 고도화된 AI 프로그램으로 발전하고 있다. 앞으로 지금보다 훨씬 고도화된 AI 기술을 적용하면 수백억 개가 넘는 각종 데이터를 실시간으로 분석해 범죄 가능 지역을 예측하고 미리 순찰함으로써 범죄율을 낮출 수 있게 될 것이다.

그런데 여기서 AI 시대에 일어날 전면적인 업무 수행 방식 변화의 일단을 볼 수 있다. 이전에는 이러한 범죄 예측 프로그램이 정보를 제공하며 경찰의 판단을 도와주는 정도였다면, 이후로는 범죄 예측 프로그램이 시키는 대로 하게 될 것이다. AI의 명령에 따라 A 경찰은 B 구역을 오전 3시에 순찰하고, C 경찰은 D 구역을 오전 4시에 순찰하는 식이다. 이렇게 되면 사람은 더는 일의 주체가 아니고, 고도화된 AI 시스템의 지시를 받아 그 수족처럼 작동하게 된다. 이런 일은 경찰의 순찰 업무뿐만 아니라 다른 영역으로도 확산될 것이다.

AI로 인한 국가 간 격차 심화

산업혁명 시대에는 생산 수단을 보유한 국가와 그렇지 못한 국가가 있었다. 그 격차가 커지면서 생산 수단을 보유한 국가는 그렇지 못한 여러 국가를 식민지로 삼았다. 앞으로 세상은 AI 코어 엔진을 보유한 국가와 그렇지 못한 국가로 나뉠 것이고, AI의 경쟁력은 이전과는 비교할 수 없는 격차를 만들어낼 것이다. AI 여부에 따라 각국은 제조, 유통, 국방, 의료, 법률 등 모든 영역에서 극명하게 차이를 겪을 수밖에 없으며 그 격차는 점점 더 벌어지게 된다.

AI와 차별 문제

세계적인 검색 사이트에서 'beautiful(아름다운)'을 입력하면 온통 백인 여성의 모습만 나오고, 'professor(교수)'라는 단어를 입력하면 주로 남성이 등장하는 등 인종과 성 차별적인 요소들이 디지털 검색 결과물에 포함되어 논란을 일으킨 적이 있다. 만약에 AI에도 이런 차별적인 요소가 포함된다면 그 여파는 훨씬 클 것이다. 물론 어느 정도의 차별은 비즈니스 차원에서 용인되어왔다. 예를 들어, 비즈니스 콜센터에 전화할 때 큰 잠재 수익이 예상되는 고객이라면 상담원과 바로 연결되지만 그렇지 않으면 상담원에게 연결되기까지 대기 시간이 길어지는 일은 지금도 일어난다. 하지만 응급 상황에서도 이러한 차별적 대응이 이루어진다면 그것은 큰 문제가 될 것이다.

AI로 인한 위험, 어떻게 대비할 것인가

AI로 인한 X이벤트급 재난을 막는 가장 확실한 방법은 AI를 사용하지 않는 것이다. 그러나 이것은 전기를 사용하지 않고 사는 것보다 더 힘든 일이 되어가고 있다. AI의 효율성과 가치는 앞으로 더 커질 것이며 우리 삶의 모든 영역에 밀접하게 관여할 것이다. 이럴수록 우리는 AI가 사람을 도와주는 보조적 수단이라는 것을 잊지 말고, 인간 중심 AI human-centered AI로 발전할 수 있도록 더 많은 사회적 관심과 노력을 기울여야 한다. AI의 복잡성이 인간의 개입 능력을 초과하기 이전 단계인 지금이 그러한 위험 요인을 대비할 수 있는 적기다.

공정한 알고리즘을 위한 사회적 통제

앞서 AI의 위험 요인 중 하나로 차별 문제를 살펴보았다. 대응 방안은 의외로 간단하다. 공정성을 확보하는 것이다. 현재 사회는 AI 사회로 접어드는 초입 단계라고 볼 수 있다. 따라서 AI의 속도와 효율, 확장성도 중요하지만, 인종, 성차별 등 편견이 개입되지 않도록 처음부터 공정하게 설계하는 것이 더 중요하다. 데이터 선택 기준, 가중치 비율 등을 정할 때 지속적인 감시를 통해 AI에 대한 사회적 통제가 이뤄져야 한다.

사회적 통제권은 정보 공개 청구와 같은 방법을 생각해볼 수 있다. AI의 알고리즘 작동 기준과 가중치, 데이터 선택 등에 관해 국회나 사회단체 등에서 공개를 요청하는 경우 투명하게 공개하고 사회 불평등이나 차별, 편견의 요소가 있다면 보완할 수 있는 체계를 구축해야 한다. 이러한 투명성transparency을 확보하기 위해 가장 중요한 부분이 AI 알고리즘을 운용하는 기관의 설명성explainability을 확보하는 것이다. 보통 AI

를 가동하는 업체는 이런 요구에 응하는 것을 영업 비밀이나 기술 노하우를 공개하는 것처럼 여겨 민감하게 대응하는 경향이 있다. 그러나 지능형 알고리즘을 사용하는 기업, 기관, 국가 등은 알고리즘의 작동 방식을 일반인들도 알 수 있도록 쉽게 설명할 의무가 있다. 어디에 가중치를 두고 있으며, 어떤 데이터를 선택해 운용하는지 사회가 알 수 있도록 해야 한다.

이런 점에서 2019년 발의된 미국의 '알고리즘 책무성 법안Algorithmic Accountability Act'은 많은 것을 시사한다. 이 법안은 불투명한 알고리즘에 대한 대책으로 한국의 공정거래위원회에 해당하는 연방거래위원회FTC에 다음과 같은 권한을 부여하고 있다.

1. 알고리즘의 작동 원리를 모니터링한다.
2. 알고리즘이 공정하고 정확하게 작동하는지 감시한다.
3. 연간 매출이 5,000만 달러 이상인 기업에 적용한다.
4. 지식재산권 문제가 관련되어 있어도 공개적으로 감사를 진행할 수 있도록 한다.

우리는 자동차를 주기적으로 정비하고 검사한다. 정비 소홀로 사고가 발생하지 않도록 하기 위함이다. AI의 알고리즘에 문제가 없는지를 살펴보는 것도 같은 맥락이다. AI의 알고리즘에 전문가만이 관여할 수 있다고 생각하는 사람도 적지 않지만, 정말로 그렇다면 앞으로 인간의 삶에 많은 영향을 미치게 될 AI 알고리즘에 오직 소수의 의견만 반영될 수도 있다. 따라서 다른 사회적 이슈에 대처할 때처럼 AI 알고리즘 가동에 대해서도 전 사회가 관심을 쏟아야 한다.

최상위 권한은 인간에게 준다

가까운 미래에는 기능별 AI를 통제하고 관리하는 '메타 AI'가 등장할 가능성이 크다. 예를 들어 헬스케어에 특화된 AI가 기능별 AI라면, 메타 AI는 여러 분야나 기능별 AI를 더 높은 차원에서 인식하고 관리·통제하는 능력을 지닌 최상위·초거대 AI를 말한다. 모든 것을 관장하는 메타 AI가 더 효율적이고 편할 수도 있다. 그러나 그 부작용도 생각해봐야 한다. 메타 AI에 문제가 생기면 관련된 모든 기능별 AI에도 연쇄적으로 문제가 발생할 수 있고, 특히 메타 AI가 해킹되거나 오작동할 시에는 더 치명적인 결과로 이어질 수 있다. 따라서 우리는 메타 AI에 최상위 권한을 부여하는 방식보다는 인간의 통제를 받는 기능별 AI를 발전시키는 방향을 고민해야 한다.

단절 구역 설정하기

코로나19 사태에는 인간의 자연 침범이나 불량한 위생 환경 등 여러 요인이 있었지만, 이것이 팬데믹으로 발전하게 된 것은 '연결'의 문제 때문이었다. AI의 경우에도 네트워크에 연결된 특정 노드에서 치명적 바이러스가 발생하거나 오작동이 일어나면 순식간에 전 세계에 그 여파가 미칠 수 있다. 따라서 매년 연결에서 완전히 단절된 지역을 확보하는 방안을 생각해볼 필요가 있다. 가령, 매년 특정 지역을 단절 구역으로 선정해 AI 시스템과 접속되지 않은 상태로 가동하다가, 1년 후 다시 네트워크에 연결해 복제와 동기화를 진행하는 식이다. 이렇게 하면 AI 시스템에 X이벤트급 사고가 발생하더라도 복구 거점을 확보할 수 있다.

백업의 중요성

재앙이 될 사안에 대해서는 최소 비용, 최대 효율의 논리가 아닌 과잉과 잉여의 논리로 대응하는 것이 적절하다. 예를 들어, 인터넷이 인공위성을 통해 전 지구적 네트워크로 연결되더라도 기존의 연결 방식인 해저 케이블이나 네트워크 회선 등을 폐기하지 않고 백업 네트워크로 보존할 필요가 있다. 그렇게 하면 만에 하나 인공위성이 X이벤트급 재난으로 파괴되거나 가동을 멈춘다 해도 최후의 방법으로 과거 방식의 백업 네트워크를 가동할 수 있기 때문이다. 이처럼 인류의 삶에 결정적인 영향을 미치는 부분에서는 사회적 비용이 엄청나게 많이 들더라도 과잉과 잉여의 논리로 2차, 3차의 백업 방안을 마련해두어야 한다.

AI 역시 오작동이나 심각한 해킹으로 인해 초기화하거나 폐기해야 하는 상황이 생길 수 있다. 이런 상황에 대비해 시스템 체계를 모두 백업해둔다면 피해를 최소화하고 대응할 시간을 확보할 수 있을 것이다. 파괴되고 무너지고 잃은 후가 아니라 그 전에 대비해야 한다.

유전자 가위 기술에 의한 차별적 미래 사회

아래 시나리오를 한번 상상해보자. 윤리 의식이나 법적 통제를 배제한다면 이런 상상이 현실이 될 수 있을 만큼 생명공학 기술이 빠르게 발전하고 있다. 현재 우리 통념으로는 불편한 일이 미래에는 빈번하게 일어날지도 모른다.

가상 시나리오 1: 죽은 아내의 유산

A는 7년의 연애 끝에 결혼했다. 그러나 신혼의 단꿈이 가시기도 전에 아내가 폐암 말기 판정을 받았고 오래지 않아 세상을 떠났다. 홀로 남겨진 A는 재혼을 생각해보라는 주위의 권유에도 불구하고 아내와의 기억을 간직한 채 독신으로 살기로 했다. 그러던 중 아내의 제대혈(탯줄 혈액)이 병원에 보관되어 있음을 알게 되었고, 제대혈을 난자로 분화하는 기술을 통해 난자를 만든 뒤 자신의 정자와 시험관 수정을 했다. 그 후 아내가 갖고 있던 발암 유전자를

유전자 가위 기술로 교정한 후 대리모에게 수정란을 이식해 건강한 딸 쌍둥이를 갖게 되었다.

가상 시나리오 2: 여성 없는 출산

A의 직장 동료인 B는 결혼할 마음이 없어 독신을 고집해왔지만, 자식을 갖고 싶은 마음이 생겼다. 그는 주위의 입양 권고 대신 자신의 체세포를 통해 인공 정자와 인공 난자를 만들었고 이를 시험관에서 수정시켰다. B의 유전자 분석 결과 당뇨병과 뇌졸중 유발 유전형이 발견되어 수정란 상태에서 유전자 교정 시술을 했다. 교정된 수정란을 8세포기까지 진행한 후 XY 염색체를 가진 수정란만을 골랐고, 대리모조차 원하지 않던 터라 인공부화 장치를 이용해 남자아이를 얻었다. 아이의 외모는 아빠와 똑같았다.

호모 사피엔스, 자연을 재설계하다

인간은 과학기술의 발전에 따라 세 단계로 분류될 수 있다. 첫 단계인 호모 사피엔스 1.0은 초기 인류의 형태다. 다른 동물과 같이 자연의 지배를 받으며 살아가고, 의식주 해결을 포함한 생활사가 자연과 밀접하게 연관되어 있다. 수렵이나 채집을 통해 먹을 것을 해결했지만 가끔 맹수들의 습격으로 목숨을 잃기도 했다. 전염병이 돌면 많은 희생자가 나왔다. 자연의 법칙을 어기지 않고 살아가는 것이 호모 사피엔스 1.0의 최고의 선택이었다.

1.0 시대가 '자연 순화형'이라면 2.0 시대는 '자연 극복형'이다. 과학기술을 바탕으로 거리, 시간, 온도 등 다양한 물리 법칙을 극복하게 되

었고, 약물과 백신의 개발로 많은 질병도 극복했다. 이제 과학기술, 특히 생명공학 기술이 극도로 발전하면서 호모 사피엔스 3.0 시대를 예고했다. 호모 사피엔스 3.0 시대는 한마디로 '자연 재설계형'이다.

유전자를 마음대로 편집하는 유전자 가위 기술

인간을 포함한 생물의 다양성은 어떻게 유지될까? 개체의 다양성 유지에는 생물학적 원천이 있다. 첫째는 DNA의 변이다. 속도가 느리고 미약한 수준이기는 하지만, 생물체의 유전자 서열은 조금씩 변한다. DNA의 복제가 완전하지 않기 때문이다. 이는 오랜 기간 생물체의 다양성과 진화의 원동력이 되어왔다. 둘째는 모계와 부계로부터 받은 염색체의 자연적 재조합이다. 후손의 염색체는 엄마나 아빠의 것 그대로 유지되지 않고 일정 부분 재조합되어 형질을 바꾸는 토대가 된다. 염색체의 자연적 재조합은 무작위적이며 예측할 수 없다.

반면, 유전자 가위• 기술을 이용하면 매우 정교하게, 그리고 의도한 대로 유전자를 재조합할 수 있다. 국소적인 질환 유전자를 바꿀 수도 있고, 큰 범위의 유전자군 조합도 바꿀 수 있다. 인위적으로 유전자를 설계하고 편집할 수 있다는 뜻이다. 다만 이 과정에서 유전자 가위 기술의 고도화와 더불어 의도된 조합으로 나타나는 형질을 예측하는 분석 기술이 필요하다. 유전자의 형질 또는 질병과의 상관관계를 깊이 파악하게 되면 우리는 유전자 설계에 필요한 정밀한 지도를 얻게 될 것이다.

• 특정 염기 서열을 인지해 해당 부위의 DNA를 절단하는 제한 효소.

유전자 가위 기술은 어떤 변화를 가져올 것인가

바이러스와 질병: 사후 치료에서 자가 치료로

바이러스는 동물, 식물, 미생물 가릴 것 없이 자신의 유전자를 복제해 줄 리보솜ribosome을 장착한 세포에는 모두 침범해왔다. 이에 세포들은 바이러스에 대응할 수 있는 나름의 체계를 갖게 되었는데, 흥미로운 체계 하나가 미생물의 '크리스퍼CRISPR 시스템'이다. 크리스퍼 시스템을 갖춘 미생물은 카스Cas라는 유전자 도구를 이용해서 침입해 들어온 바이러스 단편을 잘라 크리스퍼 영역에 삽입함으로써 외부 DNA의 출입 기록을 저장해둔다. 그리고 그 바이러스가 다시 침범했을 때 기록장치에서 '추적용 RNACRISPR RNA'를 생산해 Cas단백질과 함께 바이러스 유전자를 절단해버린다. 현재 지구상에 존재하는 많은 미생물이 크리스퍼 시스템을 갖추고 있다. 인간과 달리 미생물은 생활사가 짧으면서 오랫동안 진화를 거쳐왔기에 이렇게 멋진 도구를 실험할 기회를 충분히 가질 수 있었다.

인간도 면역 시스템을 갖고 있지만, 항체 기반의 면역 체계가 가동하기까지 시간이 걸린다. 또 급격히 증식하는 바이러스로 면역 체계가 무력화되기도 한다. 게다가 광범위한 박테리아 치료제인 항생제가 개발된 것과 달리 바이러스 치료제는 상대적으로 개발이 더디다. 현재 개발된 항바이러스제는 간염, 에이즈 등을 치료하는 몇 종류에 불과하다. 그러나 만약 위와 같은 크리스퍼 시스템을 장착한다면 다양한 바이러스의 공격에 맞설 수 있는 효과적인 면역 체계를 갖추게 되는 셈이다.

이런 점에서 치료제나 백신 개발 능력을 넘어서는 수준의 바이러스 위협이 계속된다면 인류를 상대로 한 크리스퍼의 설계적 삽입이 명분

을 갖게 될지도 모른다. 이를 통해 바이러스 감염 후의 치료적 접근이 아니라 감염이 되면 바로 크리스퍼가 작동해 선제적 치료가 이루어지게 되는 것이다. 또 바이러스 감염뿐 아니라 유전자 변이를 감지해 교정함으로써 암과 같은 질환을 자가 치료하는 기능도 수행하게 된다. 이러한 유전자 가위 도구는 '베이스 에디팅base editing'이나 '프라임 에디팅prime editing' 기술로 이미 실현되었다.

치료: 수술에서 생체 조직 교환으로

유전자 가위 기술과 줄기세포 기술의 결합은 다양한 치료를 가능하게 할 것이다. 특정 장기가 질병이나 유전적 원인에 의해 손상된 경우에도 체세포로부터 해당 장기로 분화시키는 기술이 더 발전한다면 치료할 수 있다. 필요할 경우 유전자 교정을 선행해 분화시킨 세포가 3D 프린팅, 오가노이드(미니 장기), 이종 장기(동물 장기 활용) 기술을 통해 완전한 조직으로 다시 태어나 손상된 장기를 대체하게 될 것이다. 이는 현재 제한적으로 행해지는 동종 장기 이식의 여러 문제를 해결해줄 전망이다.

수정: 자연 수정에서 배아 유도로

2021년 〈네이처〉 지는 재프로그래밍과 세포 배양 기술을 결합해 피부 세포로부터 배반포•를 형성하는 기술을 소개했다. 사람에게서 배아를 추출하지 않고 실험실에서 불임 등의 연구를 수행할 방법을 확보하게 된 것이다. 배반포 단계에서 추출한 줄기세포는 심장, 근육, 신경 등

• 포유류의 초기 발생에서, 하나의 세포로 시작된 수정란이 세포 분열을 통해서 여러 개의 세포로 이루어진 것.

의 세포로 다시 분화시킨 후 심장병, 뇌 질환, 당뇨 등의 세포 치료제로 활용될 수 있다. 그런데 배반포 그 자체를 개체로 발달시킨다면 인간 복제로 이어질 수도 있다. 핵 치환보다 더 간편하게 인간을 복제할 수 있는 기술을 인간이 손에 쥐게 되는 것이다.

탄생: 임신과 출산에서 인공 부화로

2014년 영국 과학자들은 피부 세포를 이용해 초기 단계의 인공 정자와 인공 난자를 만드는 데 성공한 바 있다. 피부 조직을 줄기세포로 재프로그래밍하고 이를 난자와 정자의 초기 세포로 발전시킴으로써 자연적 수정 과정을 거치지 않고 시험관에서 수정할 수 있는 기술적 가능성을 제기한 것이다. 2016년 중국과학원 연구팀은 실험실에서 만든 인공 정자를 난자에 주입해 태어난 건강한 쥐를 선보였다. 2021년 이스라엘의 와이즈만연구소는 시험관에서 쥐의 배아를 6일간 배양하는 데 성공했다. 쥐의 배아 발달 기간이 약 20일인 점을 고려하면 약 3분의 1을 시험관에서 발달시킨 셈이며, 그들은 이 기간에 특정 장기들이 만들어진 것을 확인했다. 향후 배양 조건 기술이 확립된다면 상상 속에서나 그려보던 고등 동물의 인공 부화도 가능해질 것이다.

죽음: 매장과 화장에서 채취와 보관으로

현재 대부분의 나라가 사체를 매장이나 화장의 형태로 처리한다. 나라마다 환생, 부활, 영혼 불멸과 같은 신앙과 믿음이 존재하더라도, 신체는 한번 죽음으로써 종결된다는 것이 지금까지의 통념이다. 하지만 과학기술의 발전으로 생전에 채취한 세포를 적절한 냉동 보관 장치에 보존함으로써 먼 미래에 유전 지도의 밑그림을 제공하는 일이 가능해

졌다. 특히 신체적·예술적 능력이 뛰어나거나 지적 능력이 탁월한 사람이라면 사후 그 신체에서 특정 세포를 채취하거나 보관하고자 하는 욕망과 필요성이 대폭 커질 것이다. 냉동 보존제, 보존 장치, 보존 후 채취 기술 등의 발달로 체세포 및 줄기세포의 보존 사례가 급격하게 확산할 수도 있다.

유전자 편집 기술의 문제들

자연 극복형의 호모 사피엔스 2.0을 넘어 3.0 시대로 질주하면서, 그동안 SF 만화나 영화에 등장했던 초현실적인 일들이 우리 앞의 현실이 될지도 모른다. 그러나 이런 기술들이 곧바로 유토피아로의 이행을 의미하지는 않는다. 예측되는 문제들을 짚어본다.

우생학과 인종주의

플라톤의《국가론》에서 선택적 남녀의 결합을 통해 인간의 종족 개선이 가능하다는 개념이 나오는 것을 보면, 우생학에 대한 역사는 매우 오래되었다. 본격적인 우생학은 찰스 다윈의 사촌인 프랜시스 골턴Francis Galton의 '인간 진화론'에 뿌리를 두고 있다. 당시 동식물의 유전학적 개량을 경험한 유전학자들을 중심으로 인간의 개량도 유전학적으로 가능하다는 신념이 공유되었다.

문제는 우생학적 사고방식이 사회·정치적 상황과 맞물렸을 때 인종청소나 홀로코스트와 같은 암울한 역사로 이어질 수 있다는 점이다. '디자이너 베이비designer baby' 기술이 소수의 국가나 집단의 전유물이 되었

을 때 심각한 인종주의로 흐를 수 있다는 점 또한 경계하지 않을 수 없다. 이러한 우생학이 우리의 현실과 동떨어진 이야기로 들릴 수도 있다. 그러나 현재 우리에게도 우생학적 사고의 그림자는 드리워져 있다. 지금 이 순간에도 산전 유전자 검사를 통한 태아 선별법이라든지, 배아 유전자 검사를 통한 배아 선별법을 통해 유전 질환의 예방이라는 명목으로 누군가가 우생학적 행동을 하고 있을지 모른다.

우월 유전자와 사회적 계층화

유전자 가위 기술에 의한 디자이너 베이비 출산은 사회 구성원을 서열화할 수 있다는 점에서 위협적이다. 특히 부의 편재와 계층화를 심화할 수 있다. 고도의 기술에 접근 가능한 상류층과 일반인의 우열 관계가 형성되면 심각한 불평등이 야기될 것이다. 우월 유전자의 존재가 증명되고 유전자의 개량이 허용된다면 우월 유전자의 대물림과 더불어 직업이나 사회적 지위의 차별은 자명해질 수밖에 없다.

유전자의 획일화

인간의 능력을 개선하는 유전자가 입증된다면, 결국 특정 유전자가 선호되면서 전체 인류의 유전자형이 특정 형태로 수렴할 위험도 존재한다. 이는 인간 다양성의 가치를 훼손하고, 장기적인 관점에서 인류의 멸종을 부를 수 있다. 생물체의 항구성은 바로 다양성에 기초하기 때문이다. 다양성의 근간이 되는 유전자 변이는 무작위적인 방향으로 발생한다. 다만 자연이라는 거대한 힘이 가장 적합한 형태로 이끌 뿐이다.

관련된 예시를 최근의 역학적 연구에서 찾아볼 수 있다. 겸상적혈구빈혈증은 유전자 이상으로 악성 빈혈이 생기는 심각한 유전 질환이지

만, 독특하게도 이 유전자형은 말라리아에 저항성을 갖는다. 말라리아가 득세하는 환경에서는 오히려 겸상적혈구빈혈증 유전자형을 가진 사람이 살아남을 가능성이 더 큰 것이다. 유전자형과 환경과의 상호 관계는 이처럼 쉽게 전체를 파악하기 어려운 면이 있다.

인간 존엄성의 훼손

치료적 관점을 넘어 강화 목적으로 시행되는 생식 세포 유전자 편집은 인간의 가치와 존엄성을 훼손한다. 인간의 유전자와 특정 형질은 임의로 자르고 붙이는 교정의 대상이 아니라 인간이 원초적으로 지닌 존엄이며, 다른 유전형을 지닌 개인들과 서로 어우러져 살아가는 것은 그 자체로 가치 있는 일이다. 특히 치료적 관점에서 유전자 교정을 하더라도 세심한 주의를 기울일 필요가 있다. 가령, 우리는 청각 장애를 심각한 장애로 규정하고 교정을 통해 청력을 회복해야 한다고 생각하지만, 반대로 청각을 회복했을 때의 괴로움과 불편함을 호소하는 예도 적지 않다. 즉, 어떤 부분에서는 장애나 질병도 상대적일 수 있음을 인식하고 늘 소통과 합의의 관점에서 접근해야 한다.

유전자 편집 기술의 위험에 대한 사회적 대응

모든 과학과 기술은 동전의 양면처럼 인간에게 유용한 도구가 될 수도 있지만 동시에 파괴적 성격을 띤 도구가 될 수도 있다. 따라서 자연 재설계적 기술을 겸허히 활용할 수 있는 지혜와 타협의 자세를 갖추어야 한다. 그럴 때 우리는 유전자 재설계 도구를 통해 악성 질병을 치료하고

동식물과 인류가 조화롭게 살아가는 환경을 만들 수 있을 것이다.

윤리적 장치의 확보

무분별한 유전자 교정에 대한 대비의 핵심은 윤리적 장치의 확보에 있다. 국내에서도 생명공학 연구와 관련해 생명윤리를 지키도록 가이드라인을 정해 명시하고 있으며, 대다수 연구 저널도 연구 윤리 및 생명윤리의 준수를 요구하고 있다. 중국 허젠쿠이 박사의 인간 배아 편집 사례는 연구 과정에서 기관생명윤리위원회IRB 심사와 동료 리뷰 시스템이 제대로 작동하지 않았다는 점에서 법적인 문제와 별개로 연구 윤리의 실천과 관리의 중요성을 시사한다.•

관련 법 제정

유전자 편집에 관한 법적·제도적 장치의 범위와 깊이는 나라마다 다르다. 일부 국가에는 관련법이 모호하거나 아예 없다. 우리나라는 '생명윤리및안전에관한법률(일명 생명윤리법)' 제47조 제3항을 통해 생식 세포, 배아, 수정란의 유전자 편집 연구 자체를 엄격히 금지하고 있다. 이에 대해 유용한 도구가 될 수 있는 유전자 편집 기술 활용을 과도하게 규제하고 있다는 의견도 있다. 불임·발달 장애와 같은 연구 및 치료를 위해 배아나 생식 세포의 유전자 교정 연구를 허용해야 한다는 이유에서다. 다만, 그 이후의 과정을 어떻게 통제할 것인가 하는 문제는 여전히 남는다. 따라서 명확한 윤리 체계와 가이드라인이 세워질 때까지는

• 2018년 허젠쿠이 박사는 유전자 편집 기술을 이용해 인간면역결핍바이러스HIV에 면역이 있는 아기를 탄생시켰다고 발표해 생명윤리에 대한 논란을 빚었다.

지금과 같은 엄격함을 한시적으로나마 유지하는 것이 바람직해보인다.

기술에 대한 폭넓은 이해를 제공하는 교육

향후 인간 유전자 교정에 관한 사회적 이슈는 점점 더 크게 부각될 것이다. 따라서 미래세대는 특히 인간 유전자 교정이 갖는 의미와 파급력 등에 대해 기술적·사회적·인본주의적 관점에서 폭넓게 이해할 필요가 있다. 이를 위해 우선 현행 교육 과정을 보완해야 한다. 과학 수업의 본래 목적인 기술적 이해를 넘어 기술이 갖는 사회·인문학적 영향에 대해 인식하고 평가할 수 있는 교육적 체계가 갖추어져야 한다.

끝없는 기술 개량과 개발

매우 발전된 크리스퍼 유전자 가위 기술에도 여전히 약점이 있다. 첫째, 원하지 않는 유전자의 편집이 일어나는 '오프타깃off-target'의 문제다. 둘째, 생식 세포 교정에서 '모자이크'가 발생할 수 있다. 수정란에 주입한 유전자 가위가 일부의 세포만 교정하는 현상이다. 이는 유전자 가위 기술만의 문제는 아니므로 다양한 해결 기술들이 함께 개발되어야 한다.

하나의 유전자는 다양한 형질과 관련이 있을 수 있다. 따라서 한 유전자를 교정했을 때 발생할 수 있는 결과에 대해 더 심층적이고 폭넓은 이해가 필요하다.

인간 뇌와 AI 결합의 가능성과 위험

인류가 인공지능AI의 잠재적 위험에 대처하는 방법을 익히지 못한다면 AI는 인류 문명사에서 최악의 재앙이 될 수 있다고 천재 물리학자 스티븐 호킹은 일찍이 경고했다. 그의 말대로라면 언젠가 AI는 인간의 능력을 넘어설 것이고, 결국 인간을 대체할 수도 있다. 그렇게 AI로 무장한 로봇이 인간을 지배하게 되는 미래를 맞이하면 인간은 어떻게 될까? 테슬라 CEO 일론 머스크의 우려처럼, 인간보다 똑똑해진 슈퍼 AI가 인간을 공격하고, 인간은 AI의 "애완 고양이house cat"가 되는 미래는 어떠할까?

이런 미래가 너무 암울하다면 다른 미래를 상상해보자. 힘들여 공부하지 않아도 영어 시험을 볼 때는 언어 영역인 뇌의 좌측 베르니케 영역에 삽입한 마이크로칩에 전류를 흐르게 해서 독해 능력을 높인다. 또 수학 문제를 풀 때는 두정엽에 삽입한 마이크로칩에 전류를 흐르게 해

탁월한 수학 계산 능력을 발휘한다. 그뿐만 아니다. 베르나르 베르베르의 소설《뇌》(열린책들, 2006)에 나오듯이 식물인간의 뇌에 전극을 이식해 컴퓨터를 조작할 수 있게 하거나, 〈매트릭스〉에서처럼 인간의 뇌와 AI 컴퓨터를 합친다.• 이런 미래가 단지 공상에 불과할까? 일론 머스크는 2021년 4월 원숭이가 말 그대로 텔레파시만으로 비디오 게임을 하는 영상을 공개했다. 이 원숭이의 뇌에는 컴퓨터 칩이 이식되어 있었다. 이번 실험은 물론 인간 두뇌에 칩을 이식하기 전 단계의 시도였다. 이렇게 텔레파시로 커뮤니케이션하고 인간의 지능이 지금보다 월등히 높아지는 '지능 증폭 사회'가 오면 인간은 더 행복해질까?

인간의 뇌에 던진 도전장

일론 머스크의 뇌–컴퓨터 인터페이스

'괴짜들의 왕'이라는 별명처럼 테슬라의 CEO 일론 머스크는 온갖 돌출 행동으로 반감을 불러일으키기도 한다. 하지만 그가 여전히 실리콘밸리의 혁신을 주도하는, 세계에서 가장 영향력이 큰 인물 중 한 명이라는 데는 반박의 여지가 없다. 그런 머스크가 큰 관심을 기울이는 분야가 인간의 뇌와 컴퓨터를 연결하는 뇌-컴퓨터 인터페이스brain-computer interface다.

그가 2017년에 설립한 회사의 이름 '뉴럴링크NeuraLink'는 인간의 뇌

• 주인공 네오의 머리에 기다란 바늘 형태의 전극을 꽂고 뇌에 무술 프로그램을 업로드하자 네오는 곧바로 무술 고수가 된다.

신경계를 무언가와 연결한다는 뜻이다. 뉴럴링크가 연결하려는 대상은 여러 가지다. 컴퓨터와 연결하면 생각만으로 마우스 커서를 제어할 수 있고, 로봇 팔과 연결하면 생각만으로 물건을 집어 올리거나 그림을 그릴 수 있다. 하지만 뉴럴링크의 궁극적인 목표는 인간의 뇌와 AI를 연결하는 데 있다.

일론 머스크의 계획은 뉴럴 레이스neural lace라고 불리는 액체 그물망 형태의 전극을 머릿속에 삽입해서 뇌 활동을 매우 정밀하게 읽어들이고 지식과 정보를 뇌에 주입하는 장치를 만들겠다는 것으로, 인간의 자연 지능과 인공지능을 연결함으로써 초지능을 구현하겠다는 의미다. 그는 인간 뇌의 모든 활동을 읽어들이면 우리 생각을 컴퓨터에 저장하는 것도, 반대로 뇌에 특정 지식을 주입하는 것도 가능할 것이라고 말한다. 물론 현재 기술로는 불가능한 일이다. 그렇다고 해서 앞으로도 불가능할 것이라는 추측은 섣부른 예단일 수 있다.

실제로 설립 2년이 지난 2019년 일론 머스크는 그간의 연구 결과를 발표했다. 이날 발표의 백미는 '신경 실' 기술이었다. 신경 실은 머리카락 굵기의 20분의 1에 불과한 4~6μm 굵기의 가느다란 실에 32개의 전극을 코팅한 뒤, 이 실을 뇌 표면에 바느질하듯이 박아 넣는다는 개념이다. 뉴럴링크는 이를 위해 초정밀 '바느질 로봇'을 만들었다. 이 로봇은 뇌혈관을 피해 출혈을 최소화하면서 자동으로 분당 6개의 실(총 192개 전극)을 뇌 표면에 이식하도록 설계됐다. 머스크는 쥐의 대뇌 피질 표면을 따라 '박음질'이 된 실 전극의 사진을 공개하기도 했는데, 이 전극들은 신호 증폭 기능이 있는 시스템 칩을 거쳐 USB-C 포트를 통해 외부 컴퓨터와 연결되었다. 다시 1년이 지난 2020년에 그는 '링크Link'라는 이름의 뇌-컴퓨터 접속 장치를 발표했다. 동전 크기의 작은 디바이스를

두개골 아래에 이식하고 1,024개의 신경 실을 로봇을 이용해 뇌에 심어 넣은 다음 무선으로 데이터를 송수신하는 시스템을 1년 만에 구현한 것이다. 이날의 발표를 한마디로 요약하자면 뇌와 컴퓨터, 더 나아가 뇌와 AI를 결합할 가능성을 좀 더 열었다는 것이다.

이러한 기술이 인간에게 적용된다고 해서 일론 머스크가 추구하는 '지식 업로드'나 '텔레파시'가 바로 구현될 수 있는 것은 아니다. 우리가 여전히 신경 세포가 만들어내는 암호를 거의 이해하지 못하고 있기 때문이다. 하지만 뉴럴링크 연구를 통해 정밀한 뇌 활동을 관찰하는 것이 가능해진다면, 우리는 '뇌의 언어'를 이해하는 데 보다 더 가까이 다가갈 수 있을 것이다.

삶의 희로애락을 기록하는 브레인 블랙박스

일론 머스크는 링크를 통해 인간의 뇌와 AI를 연결하겠다는 계획을 밝혔지만, 인간의 뇌는 전기적인 활동만으로 설명할 수 없는 부분이 많다. 다양한 호르몬과 신경전달물질의 변화에 따라서도 크게 영향을 받기 때문이다. 이러한 측면에서 볼 때, 브리티시 텔레콤의 전직 CTO이자 저명한 미래학자인 피터 코크레인Peter Cochrane 박사의 접근법은 색다르다. 그의 아이디어는 한 사람의 머릿속에 마이크로칩을 삽입해 그 사람의 일생을 기록한 뒤, 후대 사람들이 그 사람의 일생을 생생하게 경험할 수 있도록 하겠다는 것이다. 그가 '소울 캐처Soul Catcher'라고 이름 붙인 이 마이크로칩이 측정하는 것은 신경 신호에만 국한되지 않는다. 소울 캐쳐를 통해 뇌에서 발생하는 다양한 신경전달물질이나 도파민, 세로토닌 등의 호르몬까지 측정할 수 있다면 그 사람의 '감정'까지도 읽어낼 수 있을 것이다.

마인드 업로드의 꿈

2014년에 개봉한 영화 〈트랜센던스〉에는 주인공인 윌 박사의 마음을 컴퓨터에 업로드하는 장면이 등장한다. 어디까지나 SF 영화에 등장하는 설정일뿐 이 기술이 현실에서 가능하다는 이야기는 아니다. 하지만 인간의 뇌에 있는 모든 신경 회로망의 연결성 정보를 알아낸다면 컴퓨터 안에서 그 사람을 구현하는 것이 가능해질지도 모른다.

지구상에 존재하는 다세포 생명체 중에서 유일하게 모든 신경세포의 연결성 정보가 밝혀진 생명체는 1mm 정도의 몸길이를 가진 예쁜꼬마선충C. Elegans뿐인데, 과학자들은 이 생명체의 움직임을 컴퓨터 시뮬레이션으로 재현하는 데까지 이르렀다. 이러한 연구는 마인드 업로딩의 가능성을 시사한다. 실제로 2018년 MIT 출신의 로버트 맥킨타이어가 설립한 벤처 회사 넥톰Nectome은 알데히드-안정화 냉동 보관이라는 방법을 적용해 뇌 구조를 변형 없이 보존하는 기술을 보유하고 있다. 넥톰에 따르면 죽은 사람으로부터 분리된 뇌는 영하 122도에서 보관되며, 수백 년 동안 원래 상태를 유지할 수 있다고 한다.

넥톰의 아이디어는 간단하다. 언젠가 인간 뇌의 구조적·기능적 지도가 완성된다면 보존된 뇌로부터 그 사람의 모든 기억과 경험을 추출해서 컴퓨터에 업로드하겠다는 것이다. 아직은 실현이 어려운 일이다. 하지만 인간이 지구를 벗어나 우주를 여행하는 것이 상상에서나 가능한 일이었던 때를 떠올린다면 우리의 마음을 컴퓨터에 업로드하는 일도 결코 허황한 꿈이 아닐지도 모른다.

뇌공학 기술과 함께 생각해볼 문제들

이러한 일련의 뇌공학 기술의 도전에 대한 대중의 반응은 다양하다. 신체가 마비된 환자가 다시 걷게 될 것이라는 희망적인 반응도 있지만, 영화 〈공각기동대〉에서처럼 뇌에 브레인 칩을 삽입하고 살아가는 것에 대한 두려움과 걱정도 있다. 링크와 같은 기술이 치료를 위해서 쓰이는 것을 넘어 정상인들의 인지 능력을 높이는 인지 증강 수단으로 사용된다면 어떤 일이 생겨날까?

브레인 칩이 만들 계층 사회

대학 입시나 국가고시에서 인지 증강 기술로 기억 능력을 향상한 사람이 좋은 결과를 내게 된다면, 누구나 브레인 칩 이식을 받고 싶어 할 것이다. 그런데 브레인 칩 이식에는 상당한 비용이 따를 수밖에 없다. 결국 경제적 이유로 이식 수술을 받지 못해 인지 능력을 높이지 못한 채 경쟁에 뒤처지고 경쟁에서 소외되는 계층이 생겨날 것이다. 경제적 격차가 지능 격차로 연결되어 기술의 힘으로 증강된 지능과 그렇지 못한 지능에 따라 계층이 나뉜다면 지금과는 또 다른 끔찍한 불평등을 낳을 수도 있다. 또 인지 증강 기술을 특정 국가(선진국)만 독점하게 된다면 선진국과 후진국 사이의 빈부 격차는 더 커지고 전 지구적으로 불공정이 확대될 수도 있을 것이다.

전자두뇌 쇼핑 사회

자신의 의지와 관계없이 단지 사회에서 인정받기 위해서, 더 좋은 직장에 들어가기 위해서 두개골을 열고 마이크로칩을 삽입하는 사람들이

생길 수도 있다. 마이크로칩의 성능이 회사에 따라 차이가 있어 A 회사의 칩을 썼느냐 B 회사의 칩을 썼느냐에 따라 개개인의 능력이 결정될 수도 있을 것이다. 사람들은 생체 칩이 출시될 때마다 스마트폰을 교체하듯이 새로운 성능이 추가된 칩으로 손쉽게 교체하기 위해 메모리 슬롯 같은 것을 머리에 뚫고 다닐지도 모른다. 또 새로운 마이크로칩이 출시되면 너나 할 것 없이 재수술을 받으려 할 수도 있다.

뉴로 해킹의 위험

머릿속의 정보를 컴퓨터에 업로드하거나 컴퓨터 속 정보를 머릿속에 다운로드받는 과정에서 해커들이 개입할 위험성도 생각해볼 필요가 있다. 이미 '뉴로 해킹neuro-hacking'이라는 용어가 만들어져 있을 정도로 미래에는 심각한 문제가 될 가능성이 있는 일이다. 지금 논의되는 데이터 보안과는 전혀 다른 차원에서 '생각 보안' 문제가 발생할 것이다.

미지의 부작용

우리가 인간의 뇌에 대해서 완전히 이해하지 못한 상황에서 뇌를 조절함으로써 뜻밖의 부작용이 발생할 가능성도 있다. 세계적인 연구 성과를 내고 싶은 욕심에 전두엽에 인지 증폭을 위한 마이크로칩을 이식한 수학자가 있다고 가정해보자. 그는 뇌 전기 자극을 통해 얻은 통찰력과 향상된 수학 능력을 이용해 많은 수학 난제를 풀어내겠지만, 전두엽 기능의 과도한 활성화로 인해 감정이 메마른 인간이 되거나 정신질환에 걸릴지도 모른다.

인위적으로 인간의 지능을 높이는 것이 인류의 행복을 위해 바람직한지 사회적 논의가 필요하다. 뇌공학 분야 연구가 초래할지도 모르는 다

양한 사회적 문제에 대응하고 뇌공학의 올바른 발전 방향을 제시할 필요가 있다. 이 문제에 대해 고민하지 않는다면 인류는 스스로가 만든 기술에 의해 자신의 자유와 행복을 잃어버리는 우매한 종족이 될지도 모른다.

진짜 같은 가짜,
딥페이크의 위협

가상 시나리오: 아들에게서 걸려온 긴급 영상 통화

A에게 갑자기 모르는 번호로 영상 통화가 걸려왔다. 보이스피싱에 당했던 아픈 기억이 있어 받을 것인지 한참을 망설였지만 화상으로 전화가 오는 경우는 많지 않았다. 설마 보이스피싱 같은 일은 아니리라 생각하고 전화를 받았더니, 다행히 대학생 아들의 얼굴이 나타났다. 반가운 마음에 "아들! 웬일로 영상 통화를?" 하고 말을 꺼냈다. 그런데, 건너편 전화기의 카메라가 움직이자 아들이 병원 침대에 누워 있는 것이 보였다. A의 마음이 덜컹 내려앉은 순간 다른 남자가 아들 옆에서 다급한 목소리로 당장 수술이 급하다며 수술비를 빨리 보내지 않으면 수술이 어렵다고 소리쳤다. A는 의심할 겨를도 없이 보내준 계좌번호로 돈을 보낸 뒤 다시 전화가 오기를 기다렸다. 그런데 잠시 후 아들이 멀쩡한 모습으로 집에 들어서는 게 아닌가? 아들은 깜짝 놀라는 A를 보고 오히려 영문을 모르겠다는 표정을 지으며 자기 방으로 들어갔

다. A는 보이스피싱 때보다 더한 충격에 망연자실하고 말았다.

앞의 내용은 실제 일어난 사건은 아니지만, 조만간 우리 사회에 새로운 위협이 될 '비디오피싱video phishing'이라는 가상의 피해 사례다. 흔히 보이스피싱voice phishing으로 알려진 전기통신금융 사기는 범행 대상자에게 전화를 걸어 금융감독원이나 수사기관인 것처럼 속이고 송금을 요구하거나 특정 개인정보를 수집하는 사기 수법이다. 피해 사례 전파와 경찰청의 적극적인 노력으로 피해가 많이 줄긴 했지만, 수법이 지능화하면서 여전히 사회적 문제를 일으키고 있다. 그런데 만약 가족의 얼굴을 영상으로 보여주면서 사기 행위를 한다면 피해 가능성은 훨씬 더 커질 것이다. AI를 이용해 영상을 조작하는 딥페이크deepfake 기술은 보이스피싱을 넘어 비디오피싱이라는 새로운 범죄행위에 악용될 우려가 있다. 딥페이크는 AI 기술의 양면성을 보여주는 대표적 기술이다. 급속도로 발전하고 있는 AI 기술에 대해 안전장치가 필요하다는 의견이 목소리를 높여가고 있는 이유다.

딥페이크의 기술적 원리

딥페이크는 2017년 미국의 한 온라인 커뮤니티에 처음 등장했다. 스칼렛 요한슨 같은 유명 배우가 나오는 포르노 영상물이 올라온 것이다. 사실이라면 전 세계가 발칵 뒤집힐 사안이었다. 하지만 이 영상은 포르노 배우 얼굴을 유명 연예인으로 살짝 바꿔치기한 것이었다. 이 영상을 올린 사람이 사용한 대화명이 '딥페이크스deepfakes'였고, 이후 AI 기술을

이용해 조작한 영상을 딥페이크라 부르게 됐다.

딥페이크는 AI 핵심 기술인 머신러닝의 한 분야인 딥러닝deep learning과 가짜를 의미하는 페이크fake의 합성어로서 딥러닝을 이용해 영상 속 원본 이미지를 다른 이미지로 교묘하게 바꾸는 기술이다. 딥페이크는 추출, 학습, 병합의 세 단계로 진행된다. A의 얼굴에 B의 얼굴을 덧입힌다고 가정해보자. 이때 가장 먼저 수행하는 것은 추출 작업이다. A와 B의 사진 수천 장을 훑어가면서 필요한 이미지를 추출하는 과정이다. 꼭 사진일 필요는 없다. 간단한 동영상만 있어도 된다. 동영상은 짧은 순간도 엄청나게 많은 프레임으로 구성되어 있기 때문이다. 이렇게 이미지를 추출할 때 사용되는 AI 알고리즘을 인코더encoder라고 한다. 인코더는 A와 B 두 사람 얼굴에서 유사점을 찾아내고 학습한다.

학습을 끝낸 다음엔 디코더decoder 알고리즘이 압축 영상을 풀어서 얼굴 모양을 만드는 방법을 학습한다. 이 과정이 안면 매핑face mapping 혹은 안면 스와핑face swapping이다. 이 과정엔 디코더 두 개가 동원된다. 첫 번째 디코더는 A, 두 번째 디코더는 B의 얼굴을 복원하는 방법을 학습한다. 이런 학습 과정을 거친 뒤 엉뚱한 디코더에 얼굴 이미지를 보내준다. 이를테면 A의 압축 이미지를 B 얼굴을 학습한 디코더에 보내주는 방식이다. 그렇게 되면 B 얼굴을 학습한 디코더가 A 이미지로 B 얼굴을 복원하게 된다. 이 과정이 병합이다. 바뀐 얼굴 모습이 그럴듯하게 보일 수 있도록 영상의 모든 프레임에 대해 안면 스와핑 작업을 하게 된다.

딥페이크의 기술적 바탕이 된 것은 '생성적 적대 신경망GAN, generative adversarial network'이다. GAN은 머신러닝의 세 가지 핵심 분야 중 하나인 비지도 학습unsupervised learning을 이용한다. 2016년 이세돌 9단을 꺾

으면서 화제를 모았던 알파고는 지도 학습supervised learning과 강화 학습reinforce learning을 활용했다. 말하자면 정답을 반복적으로 알려주면서 훈련하는 방식이다. 반면 GAN에는 생성자와 판별자라는 두 가지 모델이 존재하는데, 이 두 모델은 위조지폐범과 이를 쫓는 경찰처럼 서로 경쟁하면서 결과물을 만들어낸다. 생성자는 데이터 학습을 토대로 거짓 데이터를 만들고 판별자는 생성된 데이터의 사실 여부를 판별하는 역할을 반복하면서, 생성자는 점점 더 진짜 같은 이미지를 만들고 판별자는 진짜와 가짜를 구분하는 능력을 계속 높여간다. 즉 가짜 얼굴을 만드는 상황에서 생성자는 가짜 얼굴을 만들어내고, 판별자는 진짜 얼굴처럼 보이는지를 판단하는 것이다. 생성자와 판별자 모두 무승부가 되는 시점이 학습의 마지막 단계가 되고, 그 결과 가짜지만 진짜와 같은 이미지가 도출된다.

딥페이크의 가능성과 위험

딥페이크가 음란물 양산에 동원되고 비디오피싱을 가능하게 하면서 그 피해에 대한 우려가 커지고 있다. 하지만 딥페이크가 꼭 부정적인 것만은 아니다. 영화를 비롯한 다양한 분야에 활용될 수 있기 때문이다. 실제로 2019년 넷플릭스 오리지널 시리즈로 개봉된 영화 〈아이리시 맨〉은 딥페이크 기술을 활용해 영화의 현실성을 높였다. 노년의 배우가 연기한 영상에 그 배우의 젊은 시절 영상에서 뽑아낸 데이터를 입히는 '디에이징de-aging' 기법을 활용해 배우의 젊은 시절 모습을 자연스럽게 구현했다. 딥페이크는 영화나 광고 영상 더빙에도 활용될 수 있다. 축구

스타 데이비드 베컴이 출연한 2019년 말라리아 퇴치 홍보 영상 제작 당시, 그가 중국어, 아랍어, 힌디어 등 9개 언어를 구사하는 장면은 이 기술 때문에 가능했다. 그 밖에도 뉴스를 전달하는 앵커의 이미지를 인공 인간으로 똑같이 재현하거나 역사 속 인물을 실제 모습처럼 불러오는 데에도 폭넓게 활용될 수 있다.

물론 악용되는 사례도 증가하고 있다. 앞에서 언급했듯이 딥페이크 기술로 유명 연예인들의 얼굴 사진을 도용해 음란물을 제작하는 것이 대표적이다. 이런 영상은 점점 더 정교해지는 AI 기술을 이용해 육안으로는 구별이 힘든 수준에 이르고 있다. 유명인뿐 아니라 소셜미디어에 영상을 올린 일반인들도 딥페이크 악용의 피해자가 될 수 있다. 사이버 보안 업체 센시티Sensity에 따르면, 딥페이크 영상의 90% 이상이 동의를 받지 않은 외설 영상이라고 한다. 네덜란드의 사이버 보안 연구 기업 딥트레이스Deeptrace의 2019년 조사에서는 딥페이크 영상 중 96%가 음란물이고, 이 중 25%는 한국 여성 연예인의 얼굴이 합성된 것으로 나타났다. 또 다른 악용 사례는 유명인의 모습에 목소리까지 위조한 가짜 뉴스를 확산시키고 사회 불안을 조장하는 경우다. 실제로 이러한 딥페이크 악용 영상이 인터넷에 떠돌기도 했는데, 마테오 렌치 전 이탈리아 총리가 다른 사람을 모욕하는 모습의 영상 등이 그 예다.

이처럼 딥페이크는 높은 활용도만큼 허위 정보를 제작해 사회적으로 혼란을 일으킬 수 있다. 더 큰 문제는 누구나 쉽게 딥페이크 기술에 접근할 수 있다는 점이다. 더욱 정교해진 딥페이크 기술이 더 많은 사람에게 피해를 줄 가능성이 커지는 것이다.

딥페이크의 세상에서 안전하게 살아가기

딥페이크의 부작용이 대두되면서 딥페이크 탐지 기술 개발 등 여러 대책도 나오고 있다. 미국 국방성 산하 고등연구계획국의 미디어 포렌식 프로그램이 대표적이다. 특히 딥페이크를 악용한 범죄에 대응하기 위해서는 기술적으로 딥페이크 영상을 정확하게 판독하는 시스템 개발이 필요하다. 딥페이크 판독 기술의 수준도 생성 수준만큼이나 빠르게 향상되고 있긴 하지만, 더 신속한 개발이 이뤄져야 한다. 시간적 측면에서 볼 때 판독 기술은 새로운 딥페이크 생성 방법에 이어 나오는 후행 기술이기 때문에, 그 공백기에는 범죄가 기승을 부릴 가능성이 계속 존재한다.

제도적으로도 딥페이크 기술로 만든 콘텐츠는 딥페이크 기술로 생성되었음을 반드시 표시하도록 해야 한다. 딥페이크 기술이 점점 고도화됨에 따라 앞으로는 구별이 거의 힘든 시점에 이를 것이고, 이에 따라 사용자는 자신이 가상 인간과 대화를 하는 것인지 아니면 실제 인간과 대화를 하는 것인지를 인지하고 대응할 필요가 있다. 좋은 의도에서 딥페이크 영상물을 만든다고 해도 당사자의 동의 없이 만드는 것이 적절한지에 대한 사회적 합의도 필요하다.

딥페이크와의 전쟁을 위한 입법 활동도 활발하다. 우리나라에서도 2020년 3월 성폭력처벌법 개정안에서 딥페이크 관련 처벌 규정을 추가했다. 다만 이 개정안은 제작 의뢰자나 소지자에 대한 처벌 규정이 빠져 있어 반쪽짜리라는 지적을 받고 있는데, 이에 따라 2021년 5월에는 처벌 대상에 딥페이크 소지·저장·시청을 한 경우도 포함하는 개정안이 발의되었다.

물론 이런 노력만으로는 부족하다. 딥페이크뿐 아니라 AI 기술이 의도적으로 악용되지 않도록 사회적 안전장치를 만들어야 한다. 인공지능 기술을 개발하는 개발자들에 대한 윤리성 제고도 필요하고 활용하는 기업이나 사람들 역시 윤리 지침을 이해하고 지켜야 한다. 우리 정부에서도 2020년 12월 '인공지능 윤리기준'을 마련하고 발표한 바 있다. 앞으로 개발될 모든 AI는 '인간성을 위한 인공지능AI for humanity'을 지향하고, 인간 삶의 질 향상에 이바지하도록 활용되어야 할 것이다.

인공지능 윤리기준

3대 기본 원칙

① 인간 존엄성 원칙

- 인간은 신체와 이성이 있는 생명체로 인공지능을 포함해 인간을 위해 개발된 기계제품과는 교환 불가능한 가치가 있다.
- 인공지능은 인간의 생명은 물론 정신적 및 신체적 건강에 해가 되지 않는 범위에서 개발 및 활용되어야 한다.
- 인공지능 개발 및 활용은 안전성과 견고성을 갖추어 인간에게 해가 되지 않도록 해야 한다.

② 사회의 공공선 원칙

- 공동체로서 사회는 가능한 한 많은 사람의 안녕과 행복이라는 가치를 추구한다.
- 인공지능은 지능정보사회에서 소외되기 쉬운 사회적 약자와 취약 계층의 접근성을 보장하도록 개발 및 활용되어야 한다.

- 공익 증진을 위한 인공지능 개발 및 활용은 사회적, 국가적, 나아가 글로벌 관점에서 인류의 보편적 복지를 향상시킬 수 있어야 한다.

③ 기술의 합목적성 원칙

- 인공지능 기술은 인류의 삶에 필요한 도구라는 목적과 의도에 부합되게 개발 및 활용되어야 하며 그 과정도 윤리적이어야 한다.
- 인류의 삶과 번영을 위한 인공지능 개발 및 활용을 장려해 진흥해야 한다.

10대 핵심 요건

① 인권 보장

- 인공지능의 개발과 활용은 모든 인간에게 동등하게 부여된 권리를 존중하고, 다양한 민주적 가치와 국제 인권법 등에 명시된 권리를 보장해야 한다.
- 인공지능의 개발과 활용은 인간의 권리와 자유를 침해해서는 안 된다.

② 사생활 보호

- 인공지능을 개발하고 활용하는 전 과정에서 개인의 사생활을 보호해야 한다.
- 인공지능 전 생애주기에 걸쳐 개인정보의 오용을 최소화하도록 노력해야 한다.

③ 다양성 존중

- 인공지능 개발 및 활용 전 단계에서 사용자의 다양성과 대표성을 반영

해야 하며, 성별·연령·장애·인종·종교·국가 등 개인 특성에 따른 편향과 차별을 최소화하고, 상용화된 인공지능은 모든 사람에게 공정하게 적용되어야 한다.

- 사회적 약자 및 취약 계층의 인공지능 기술 및 서비스에 대한 접근성을 보장하고, 인공지능이 주는 혜택은 특정 집단이 아닌 모든 사람에게 골고루 분배되도록 노력해야 한다.

④ 침해 금지

- 인공지능을 인간에게 직·간접적인 해를 입히는 목적으로 활용해서는 안 된다.
- 인공지능이 야기할 수 있는 위험과 부정적 결과에 대응 방안을 마련하도록 노력해야 한다.

⑤ 공공성

- 인공지능은 개인적 행복 추구뿐만 아니라 사회적 공공성 증진과 인류의 공동 이익을 위해 활용해야 한다.
- 인공지능은 긍정적 사회 변화를 이끄는 방향으로 활용되어야 한다.
- 인공지능의 순기능을 극대화하고 역기능을 최소화하기 위한 교육을 다방면으로 시행해야 한다.

⑥ 연대성

- 다양한 집단 간의 관계 연대성을 유지하고, 미래세대를 충분히 배려해 인공지능을 활용해야 한다.
- 인공지능 전 주기에 걸쳐 다양한 주체들의 공정한 참여 기회를 보장하

여야 한다.

- 윤리적 인공지능의 개발 및 활용에 국제사회가 협력하도록 노력해야 한다.

⑦ 데이터 관리

- 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용하고, 목적 외 용도로 활용하지 않아야 한다.
- 데이터 수집과 활용의 전 과정에서 데이터 편향성이 최소화되도록 데이터 품질과 위험을 관리해야 한다.

⑧ 책임성

- 인공지능 개발 및 활용 과정에서 책임 주체를 설정함으로써 발생할 수 있는 피해를 최소화하도록 노력해야 한다.
- 인공지능 설계 및 개발자, 서비스 제공자, 사용자 간의 책임 소재를 명확히 해야 한다.

⑨ 안전성

- 인공지능 개발 및 활용 전 과정에 걸쳐 잠재적 위험을 방지하고 안전을 보장할 수 있도록 노력해야 한다.
- 인공지능 활용 과정에서 명백한 오류 또는 침해가 발생할 때, 사용자가 그 작동을 제어할 수 있는 기능을 갖추도록 노력해야 한다.

⑩ 투명성

- 사회적 신뢰 형성을 위해 인공지능의 투명성과 설명 가능성을 높이고,

타 원칙과의 상충 관계를 고려해 활용 상황에 적합한 수준의 투명성과 설명 가능성을 높이려는 노력을 기울여야 한다.

- 인공지능 기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의 사항을 사전에 고지해야 한다.