AI 초지능의 위험을 어떻게 피할 것인가

오준호

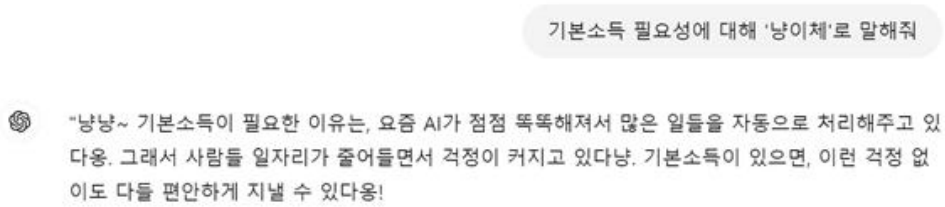
『사명이 있는 나라』 지은이.
기본소득정책연구소장

AI 혁신에 대한 개인적인 체험

얼마 전 기본소득 강연 요청을 받고, 프레젠테이션 자료 작성을 AI에게 맡겨 보기로 했다. AI 서비스는 사람들이 추천하는 감마Gamma를 택했다. 먼저 상세히 강연 원고를 써서 AI에 업로드하고 '생성' 버튼을 눌렀다. AI는 30초도 안 돼 프레젠테이션 자료를 뚝딱 만들었다. 결과물은? 기대했던 것보다 훨씬 근사했다.

AI는 원고의 중심 내용을 파악하여 깔끔하게 정리했고, 내용에 관련된 이미지를 만들어 곁들였다. AI가 만든 '기본소득을 받고 기뻐하는 사람들' 이미지가 너무 서양인 일색이기에 동양인을 포함하라고 지시하니 금방 새 이미지로 바꿔줬다. 무료 버전이라 10페이지가 최대였지만 만들어진 자료에 페이지를 추가하긴 쉬웠다. 이젠 강연 때마다 프레젠테이션 자료 만드는 스트레스는 끝이구나! (이후 다른 강연 때문에 감마 AI에 들어가니 '무료 이용'은 끝났다며 유료 구독을 권했다. 역시 공짜는 없다.)

'대생성의 시대'다. 쏟아져 나오는 생성 AI 서비스를 따라가기도 벅차다. 챗지피티 포오(4-o의 o는 '모든 것omni'를 뜻한다)는 사용자와 끊어지지 않는 음성 대화가 가능하다. 음성 대화시 말투를 바꾸게 할 수 있는데, 이런 것도 된다.



이 글을 쓰기 전에도 먼저 챗지피티와 클로드3.5 소네트에 쓰려는 주제와 몇 줄의 질문을 입력했다. AI가 제공한 개요와 샘플 원고는 좋은 참고 자료다. 요새는 자료 검색에 퍼플렉시티perplexity AI를 자주 쓴다. 질문하면 깔끔한 답변을 제공하면서, 챗지피티와 달리 답변의 출처인 기사 등을 함께 소개해 준다. AI는 영상 요약도 잘한다. '민주당 금투세 끝장토론' 영상이 너무 길어서 릴리스Lilys AI에 요약을 부탁했다. AI는 영상의 주요 구간별 핵심내용을 간추려 줬다.

최근 알게 된 놀라운 서비스는 구글이 만든 '노트북 엘엠LM'이다. 문서나 웹사이트 주소를 넣으면 AI가 그 내용으로 두 인물이 대화하는 팟캐스트를 만들어 준다. 기본소득을 주제로 내가 전에 쓴 칼럼을 올리고 팟캐스트를 생성해 보았다. 두 남녀가 "이 글은 기본소득이 필요하다고 해", "중요한 제안이야" 같은 대화를 영어로 주고받는데 어느 외국 방송에서 내 글을 가지고 진지하게 토론하는 듯한 착각마저 들었다(아직은 영어 대화만 가능하다).

나의 체험 범위는 지금 AI 혁신에서 빙산의 일각도 못 될 것이다. 이제 AI로 누구나 영상을 만들고 코딩을 한다. AI는 학생별 맞춤 학습계획과 기업별 특화된 마케팅 전략을 짜는 데 사용된다. 의료 AI는 점점 더 정확히 질병을 진단하고 효율적인 시술 방법을 제안한다. 기상 예측 AI는 기존 슈퍼컴퓨터보다 수천 배 빠르게 날씨를 예측한다(엔비디아는 자체 기상 예측 AI로 1년 뒤 허리케인의 발생을 예측할 수 있다고 한다). AI가 탑재된 휴머노이드 로봇의 대량생산도 코앞이다. 일론 머스크의 테슬라는 사람과 흡사하게 움직이는 AI 로봇 '옵티머스'를 수년 내에 대당 2,500만 원 정도 가격으로 출시할 거라고 한다. 그 가격이면 거의 모든 제조업 기업이 도입을 검토할 것이다.

AI라는 '마법의 도구'와 함께 인류는 과거에 상상할 수 없는 새로운 길에 들어섰다. 그런데 인류는 지금 어디로 가고 있을까? 이 길에 놓인 미래는 모두에게 안전하고 행복할까? 그것이 위험한 미래라면 우리는 피해 갈 수 있을까?

'초지능'은 우릴 위험한 미래로 이끌 수 있다

오픈AI 최고경영자 샘 올트만은 2024년 9월 23일 자신의 웹사이트에 "초지능이 수천 일 내 등장할 것"이라는 글을 올렸다. 그는 초지능 AI의 등장은 "높은 위험을 감수"해야 하는 일이지만 "기후를 고치고 우주 식민지를 건설하고 모든 물리학을 발견하는 놀라운 승리"를 가져올 거라고 예견했다. 그

가 초지능 AI 등장을 예견한 건 처음은 아닌데, 이전엔 AI가 인류의 위협이 될 수 있다는 우려도 내놓은 것에 비해 지금은 미래를 매우 낙관하는 것처럼 보인다.

지금의 AI를 능가하는 AGI(범용인공지능), ASI(슈퍼인공지능, 곧 초지능)가 곧 등장한다는 주장이 많다. 하지만 누구도 그게 무엇인지 딱 부러지게 정의하지 못한다. 그래서 초지능에 대한 낙관론이든 비관론이든 매우 모호하다. 나는 AI 혁신이 거듭되면 우리가 AGI나 초지능이라고 부르는 국면이 올 거라고 본다. 그러나 그 국면이 어떨지에 대해 현재의 낙관론, 비관론과는 생각이 좀 다르다.

초지능 AI에 대한 대표적인 비관론은, 의식을 갖게 된 AI가 인간을 지배하러 들 거라는 우려다. 하지만 인공지능은 아무리 성능이 발전해도 '누굴 지배하겠다'는 욕망의 감정을 갖지는 않을 것이다. 감정은 생명체의 의식 활동이고 인공지능은 기계지능이기 때문이다. 초지능 스스로 'AI 지배자'가 되기를 꿈꾸진 않을 거란 뜻이다. 하지만 낙관론이 기대하는 것처럼 AI가 인간의 번영을 이끄는 중립적 도구가 '저절로' 되지도 않을 것이다. 초지능 국면에는 새로운 위험이 등장할 가능성이 있기 때문이다.

2022년 5월, 매사추세츠 공과대학MIT과 하버드 의과대학 연구팀은 '엑스선 사진만 보고 인종을 맞추는' AI에 대한 논문을 발표했다. 이 AI는 엑스선 사진과 CT 스캔 사진만 보고 흑인과 백인을 구별해냈다고 한다. 문제는 AI가 어떻게 구별했는지 그 원리를 연구자들이 모른다는 점이다. 연구의 본래 목적은 AI가 흑인의 흉부 엑스선 사진에서 병리증상을 종종 놓치는 문제를 개선하려는 거였다. 이를 위해 AI에게 흑인과 백인의 엑스선 사진과 CT 사진을 대량 학습시키니, 나중에는 인종 정보를 따로 주지 않아도 90% 확률로 인종을 맞췄다. AI가 답을 내놓은 과정은 이른바 '블랙박스'에 감춰져 있다.

AI가 이 능력을 극대화하면, 얼굴 등 간단한 신체정보만으로도 그가 미래에 암 또는 우울증에 걸리거나, 과실이나 범죄를 저지를 가능성이 높다고 예측할 수 있다. AI가 내놓은 예측이 맞았다는 증거가 몇 차례 확인되면 AI 예측은 신뢰를 얻게 된다. 이후로 AI가 위험군으로 판정한 사람은 취업, 승진, 보험 가입 등에 불이익을 받을 수 있다.

일본 추리소설가 히가시노 게이고의 『라플라스의 마녀』에는 주변의 물리적 정보를 종합해 미래를 예측하는 천재 주인공이 등장한다. 바람의 세기나 방향, 땅의 형태나 기울기를 근거로 몇 분 뒤에 일어날 일을 정확히 예견하는 것이다. AI 기술이 닿으려는 경지가 바로 이 '라플라스의 마녀'다. 다음의 경우를 상상해 보자. 가까운 미래에 어떤 AI 기업이 대형재해 발생을 예측했다. 하지만 기업은 AI가 어떤 근거로 재해를 예측했는지 설명할 수 없었고, 정부는 이 경고를 받아들이지 않았다. 그런데 실제로 재해가 발생해 수백, 수천 명이 목숨을 잃었다고 하자. 대중은 AI의 경고를 무시한 정부에 분노하고 거리로 쏟아져 나올 것이다.

혹은 사람 표정에서 범죄 의도를 읽어내는 AI가 개발돼, 누군가를 잠재적 범죄자로 지목했다고 하자. 그를 체포할 법규가 없으니 경찰이 조치를 취하지 않았는데, AI의 경고대로 실제 범죄가 일어났다면 여론이 어떨까. 주식시장을 읽는 AI가 갑작스러운 주가 폭락을 예견했는데 정부 경제부처가 조치하지 않아 개미들이 큰돈을 잃었다면? 야당은 선거에 이기기 위해 "우리 당은 국민의 재산을 지키기 위해 AI의 결정을 전적으로 따르겠습니다"라고 하지 않을까. 이렇게 되면 자연히 모든 공공정책은 AI의

판단을 '반드시, 최대한' 참고해야 한다는 법이 나올 것이다. AI의 위상은 '초지능'으로 올라간다.

AI가 어떤 객관적인 능력치를 달성해서 초지능으로 인정받는 게 아니라, AI의 위상이 사회적으로 절대시되면 그것이 곧 초지능이다. AI의 능력이 절대시되면 이제 감히 AI의 판단을 거부할 수 없다. 그런 주장을 하면 국가 안보를 해치는 일로 여겨질 수 있다. 어떤 계기로든 AI의 판단은 무조건 따라야 한다고 사람들이 믿는 순간, 모든 영역에 AI 도입이 급물살을 탈 것이다. 사회의 모든 영역이 AI로 연결되고 AI의 판단에 사회의 운영을 맡기게 되면, 그것이 AGI이고 초지능 국면이다.

이런 상황이 오면 지금 우리가 AI에 대해 두려워하는 모든 일이 엄청난 속도로 벌어질 수 있다. AI 자동화는 필연적으로 대량 해고를 가져오고, 사회적 불평등이 빠르게 커질 것이다. AI에 의한 감시와 사회 통제가 치안의 이름으로 정당화될 수 있다. AI에 의해 직무에서 배제되고 복지제도 수급에서 탈락하는 사람들이 늘지만, 항의해보았자 기업주나 공무원은 AI의 결정이고 그 이유는 자기도 모른다고 말할 것이다.

이것은 지금의 사회적 맥락에서 AI가 발전할 때 다다를 수 있는 아주 위험한 미래다. 우리는 위험한 미래를, 모두를 위한 좋은 미래로 바꾸어야만 한다. 하지만 AI 혁신 자체를 막거나 거꾸로 돌릴 필요는 없다. AI 혁신을 가속하며 인류의 더 나은 미래를 설계할 수도 있다. 더 정확히 말하면, 사회 전체를 새롭게 디자인할 때 AI 혁신도 더 앞당길 수 있다.

'AI 커먼즈'와 기본소득으로 다른 미래를 그리자

기술 발전, 특히 AI와 관련하여 서구 좌파 이론가들 사이에 '최소주의'와 '최대주의' 입장이 대립한다. 두 입장은 기술 발전에 대한 관점 그리고 기술이 사회적 평등과 정의에 어떻게 기여해야 하는지에 대해 시각 차이가 뚜렷하다.

최소주의자들은 기술의 발전 가능성에 회의적이고, 기술이 사회문제를 자동적으로 해결하지는 못한다고 한다. 특히 AI에 의한 완전한 자동화나 인간 노동력 대체는 실제로 일어나기 어렵다고 본다. AI 기술의 현란한 간판 뒤에는 데이터 라벨링 같은 저임금 노동이 '유령노동'으로 존재하기 때문이다. 'AI 자동화'는 자본이 노동의 협상력을 떨어뜨리려고 퍼트리는 과장된 선전일 뿐이며, 좌파의 역할은 노동 보호와 고용 안정화를 위해 노력하는 것이다.

반면 최대주의자들은 기술 발전, 특히 AI 자동화가 사회적 평등을 촉진하고 인간 삶을 개선하는 커다란 잠재력을 가졌다고 본다. 이들은 기술이 노동 부담을 줄이고 사회를 풍요롭게 만들 효과적 도구이며, 나아가 사회 변혁의 무기라고 여긴다. AI 자동화는 노동시간 단축과 보편적 기본소득 도입의 계기이고, 이를 통해 사람들에게 실질적 자유를 선사할 수 있다. 또 최대주의는 기술 발전의 혜택을 공평하게 분배하기 위해 기술의 공적 소유와 민주적 통제의 필요성을 강조한다. 최대주의 입장은 가속주의 accelerationism라고도 불린다. 마르크스가 생산력과 생산관계의 모순을 밀어붙여 체제를 변혁하자고

한 것처럼, 가속주의자도 기술 발전과 사회혁신에 가속 페달을 밟아 자본주의 체제를 '돌파'하자고 한다. (최소주의와 최대주의에 대한 이상의 논의는 조정환 등이 공저한 『인공지능, 플랫폼, 노동의 미래』를 참고했다.)

그동안 한국 진보 진영의 주류적 입장은 최소주의에 가까웠다. 진보 진영 내 기본소득 반대론자들의 입장은 AI가 인간 노동을 대체하는 것은 먼 미래에나 있을 일이니 기본소득은 그때 가서 논의하자는 거였다. 지금도 진보 진영이 AI를 바라보는 일반적인 태도는 AI 혁신을 이용해 사회를 어떻게 디자인하자는 적극적인 자세보다는, AI 발전에 따른 노동의 불안정 등 부작용을 감시하자는 정도의 소극적 태도다.

자본주의 체제를 넘어서고 싶은 진보파라면 최대주의 전망을 담대하게 제출해야 하지 않을까? 물론 최대주의가 지나친 낙관주의로 빠지는 건 경계해야 한다. 지금의 사회적 조건에서 AI 혁신은 노동자에게 알고리즘에 의한 노동 통제, 저임금 따위를 받아들이라고 강요하는 '자본의 무기'가 될 가능성이 높다. 그러나 이 상황은 진보파가 AI 기술의 잠재력을 외면하고 그저 부작용에만 초점을 맞춘다면 더 악화하기만 할 것이다. 진보파는 모두에게 바람직한 미래를 위해 AI 혁신을 어떻게 활용할지 적극적 계획을 제시해서 대중의 지지를 얻어야 한다.

그러한 계획의 핵심에는 'AI 커먼즈commons'와 '보편적 기본소득'이 있어야 한다. AI의 원료는 시민들이 제공하는 데이터인 만큼 AI의 공유부적 성격이 인정되어야 한다. 또한 AI 알고리즘이 차별과 배제를 강화하지 않도록 AI에 대한 민주적 통제가 필요하다. 그리고 AI 혁신에 필요한 컴퓨팅 자원, 탄소 배출을 늘리지 않는 에너지의 공급은 국가의 대대적 투자가 있어야 확보할 수 있다. 이처럼 AI의 공유부적 성격, AI에 대한 민주적 통제, AI 혁신을 위한 공적 투자를 종합한 개념이 AI 커먼즈다. 그리고 커먼즈로서 AI를 발전시켜 얻는 이익의 일정한 몫은 마땅히 모두에게 기본소득으로 돌려줘야 한다. 보편적 기본소득은 AI 자동화를 생계 노동시간의 과감한 축소와 여가 증대로 이어지게 할 방편이자, 각자의 삶에 자유로운 기회를 꽃피게 할 수단이다. 단, 기본소득이 그러한 역할을 하려면 지급 수준이 충분히 '해방적'이어야 할 것이다.

이러한 방식으로 우리는 AI의 초지능 국면이 가져올 수 있는 위험한 미래를 피하고, '모두를 위한 AI' 시대로 나아갈 수 있다. 보편적 기본소득, 노동시간 단축 그리고 모두를 위한 AI. 이것이 우리가 생성해야 하는 더 나은 미래다.