인공지능(AI)과 신흥기술안보, 그리고 데이터안보*

김준연 소프트웨어정책연구소 산업정책연구실 책임연구원 | catchup@spri.kr
박강민 소프트웨어정책연구소 AI정책연구실 선임연구원 | gangmin.park@spri.kr

1 디지털 신기술 출현과 신흥 안보에 대한 진화적 관점

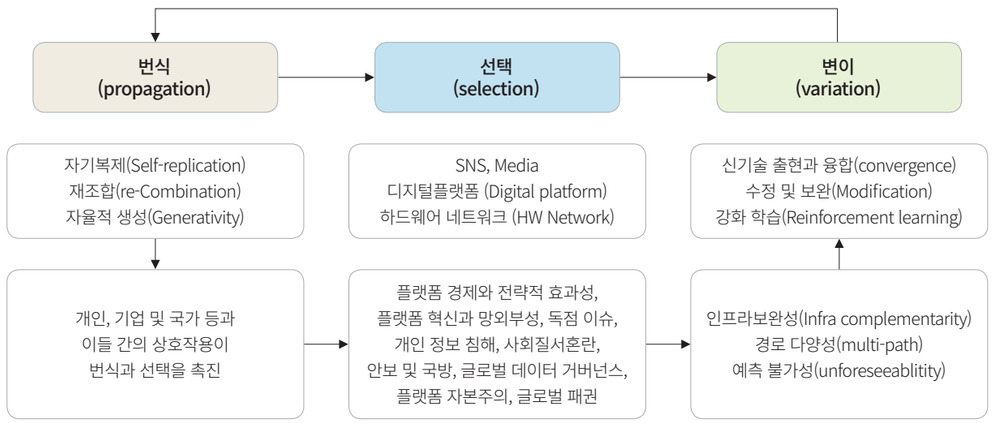
챗GPT 사용이 허가된 지 불과 20일 만에 삼성전자 반도체 부문에서 정보 유출 사고가 발생했다. 엔지니어들이 소스코드 오류 확인, 불량 설비 파악에 사용되는 코드 최적화, 녹취 전사본을 기반으로 한 회의록 작성을 위해 챗GPT를 사용했는데 이 과정에서 민감한 기업 정보가 사실상 제3자에게 넘겨져 회사의 통제를 벗어난 것이다. 최근에는 인간의 개입 없이도 목표를 달성할 때까지 끊임없이 스스로 작업을 수행하는 '오토GPT(AutoGPT)' 또는 '에이전트GPT(AgentGPT)'의 출현이 임박하면서 디지털 신기술의 출현이 촉발하는 안보위협에 대한 해석과 후과에 대한 새로운 시각이 요구되는 상황이다. 그간의 디지털 보안에 대한 접근이 주로 기술과 법학적 관점에서 진행됐다면, 이 글은 진화경제학(혹은 기술경제학; Veblen, 1989)[1]의 관점에 따라 생물학적 진화의 과정과 유사하게 경제사회공간에서도 무작위적인 변이가 시간이 지남에 따라 축적 및 번식해서 완전히 새로운 형태(종 분화)의 출현을 초래하고 대규모 변화로 우세해지는

* 제7차 사이버 국가전략포럼에서 발표한 내용을 인용

1 Veblen(1898)의 진화경제학(evolutionary economics)에 출발해 Richard Nelson, Sidney G. Winter, An Evolutionary Theory of Economic Change가 대표적 저서에 해당한다.

과정은 다시 사회적 공진화를 거치면서 인식과 제도변화를 촉발한다고 보고 있다.[2] 예컨대 AI 등 디지털 기술에 의해 조성된 생태계가 혁신을 위해 변이와 번식 및 선택의 과정을 반복하는 과정에서 새로운 유형의 공수비대칭성의 위협을 발생시키는 기제도 역시 변이 발생→번식과 증폭→선택의 피드백 과정으로 증폭되고, 이 두 가지 기제 간에 긴장과 충돌이 우리 사회에 새로운 안보 이슈를 발생시킨다고 보는 것이다. 이러한 변이와 선택의 증폭 과정이 끊임없이 병존하며 발생하는 이유는 혁신과 안보가 신기술 출현을 촉발하는 불가분의 결과물이기 때문이다. 이 글의 순서는 먼저 2장에서 AI와 데이터 생태계의 기술특성을 살펴보고, 3장에서는 미래 디지털기술의 전망을 소개하며, 새롭게 부상하고 있는 안보의 새로운 유형 몇 가지를 소개하고자 한다. 4장은 요약과 시사점이다.

[그림 1] 데이터와 인공지능 및 디지털 플랫폼 생태계의 번식과 되먹임 과정
Evolutional Propagation & Feedback Process of Data, AI and Digital Platform



2 AI와 데이터 생태계의 기술특성[3]

AI도 SW의 일부로 SW의 일반적 속성을 대부분 반영하고 있으며, 이러한 특징은 관련 생태계와 인프라에서 우위를 차지하고 있는 선발자 그룹의 전략에 유리하게 작동한다. AI의 기술적 특성이 혁신의 유형을 결정하지만, 안보라는 관점에서도 AI의 기술적 특징은 안보의 유형을 결정하는 결정 변수가 되기에 중요하다. 예컨대 결과 값의 난해한 해석과 데이터의 오염, HW 인프라 의존성으로부터 오는 안보 위협이 왜

2 Saviotti, P. P. (1996). Technological evolution, variety and the economy. Books.

3 Jacobides, M. G., Brusoni, S., & Candelon, F. (2021). The evolutionary dynamics of the artificial intelligence ecosystem. Strategy Science, 6(4), 412-435.

발생하는지에 대해서 AI의 기술적 특성은 그 실마리를 제공하기 때문이다. 또한 이러한 특성과 사회체제적 특성을 연결하면 최근 등장하는 AI 가짜뉴스, AI 여론형성 등과 같은 새로운 유형의 안보 위협도 어느 정도 해석이 가능한 것이다.

안보의 관점에서 바라본 AI의 기술적 특성들은 다음과 같다. 첫째, AI의 프렉탈 구조적 특성이다. 통상 전체 구조를 1/a로 분할했을 때, 각 부분은 전체와 통계적으로 자기 닮음이고, 그 개수가 b개일 때, 전체 구조의 프랙탈 차원이라고 한다. 프랙탈 구조의 장점은 간단한 방법으로 복잡한 구조를 만드는 데에 있다. 다만 AI의 경우, 알고리즘과 통계적 확장을 통해 프랙탈 구조를 형성함으로써 다양하고 복잡한 응용이 가능하기 때문에 결과물의 해석과 파악에 어려움이 존재할 수 있다. 실제 거대 언어모델과 같은 생성형 AI의 알고리즘 구조는 수천억 개의 파라미터(Parameter)를 가져도 프랙탈 특성에 따라 복잡한 연산이 가능한 것인데, 이 복잡한 내부의 메커니즘을 인간이 직관으로 파악하기란 불가능하다.

둘째, 데이터 의존형이라는 특징이 있다. AI 모델의 개발 과정을 보면 먼저 방대한 데이터를 확보해 학습 데이터를 구축하고, 이를 기반으로 학습하고 훈련시키면서 AI 모델을 개발하게 된다. 다만, 이러한 과정에서 학습 데이터의 일부를 추출해 내는 모형전복(Model Inversion), 데이터나 AI 모형에 대한 오염(Poisoning)[4], 데이터 착란과 오염을 넘어 AI 모형 자체에 대한 오염 공격, 입력값에 노이즈를 추가해 AI 모형이 정확한 판단을 하지 못하도록 유도하는 회피(Evasion) 공격 등이 발생할 수 있는 것이다.

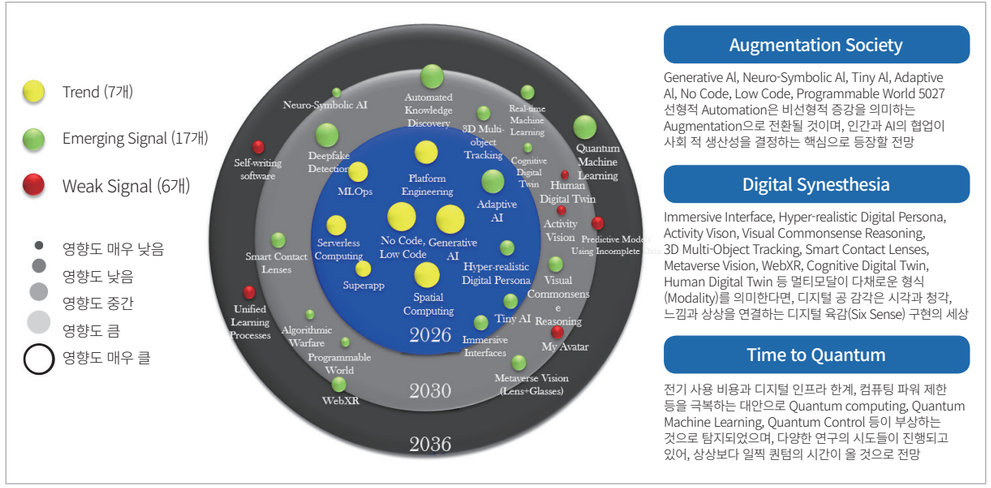
셋째, 자율적 알고리즘 특성이다. 인공지능의 자율성이란 주변 환경을 관측(Observe)하고, 판단(Orient)해서, 결심(Decide)한 후 행동(Act)하는 의사 결정 과정인 OODA 루프(Loop)상의 각 단계별로 인공지능 스스로 결정할 수 있는 능력을 의미한다. 문제는 AI가 스스로 내리는 결정을 우리가 완전히 이해하지 못한다는 것이다. 인간이 AI의 자율성에 대해 이해하기 어려워하는 현상을 "우연한 자율성(Accidental Autonomy)"이라고 설명하기도 하는데, 2019년 보잉 737 맥스의 추락 원인이 된 소프트웨어에 대해서도 당시 조종사 누구도 해당 시스템을 이해하지 못하고 있었던 사례가 대표적이다.

넷째, AI는 개방형 생태계이다. 오픈소스 소프트웨어는 소스 코드를 누구나 열람할 수 있으며, 필요에 따라 수정할 수 있다. 다만 개방형의 장점에도 불구하고 보안의 취약성도 구조적으로 발생할 개연성이 높다. 우선 많은 사용자들이 자발적으로 기여하기 때문에, 새로운 기능이나 버그 수정이 빠르게 이뤄질 수 있지만, 이 과정에서 영향력을 행사하려는 악의적 개발자의 개입 이슈가 언제든지 발생할 가능성이 있고, 대부분의 오픈소스 커뮤니티가 다른 오픈소스 커뮤니티와 의존관계에 있어서 사소한 보안 취약점이라도 발생하면 생태계 전반에 매우 큰 영향을 미칠 수 있다. AI 개발 과정 자체가 개방형의 파편화된 구조라서 원인 파악도 쉬운 일이 아니다.

다섯째, AI는 클라우드 의존적 특성이 있다. 챗GPT가 MS 클라우드 인프라에서 작동했듯이 고성능컴퓨팅(HPC)이 필수적인 AI 특성상 클라우드 기반의 서비스형 인프라(IaaS)에서 구동된다. 이를 보안의 관점에서 보면, 문제가 AI인지 아니면 클라우드 인프라에서 기인한 것인지를 명확하게 밝히기가 쉽지 않을 수 있다는 점이 있다. 이러한 상호의존성이 높은 시스템을 '복합 시스템'이라고 하는데 AI는 SW 개발과정과 구동의 인프라 모두가 외부 시스템과의 의존성이 높아서 보안 문제의 발단을 규명하는 것이 대단히 복잡해지는 측면이 있다.

3 미래 디지털기술과 신흥안보의 유형

[그림 2] 미래 디지털 신기술 트렌드 전망[5]



2024 미래 디지털기술전망(SPRi)[6]에 따르면, 2026년까지 눈여겨 볼 트렌드 기술 7개는 ▲생성AI ▲플랫폼 엔지니어링 ▲ML옵스(MLOps) ▲서버리스 컴퓨팅 ▲슈퍼앱 ▲스패셜(Spatial) 컴퓨팅 ▲노코드로코드(NCLC) 등으로 AI와 컴퓨팅 분야가 많았으며, 또한 2030년까지 유망한 이머징 시그널 기술 17개는 ▲어댑티브AI ▲하이퍼-리얼리스틱 디지털 페르소나 ▲타이니AI ▲이머시브 인터페이스 ▲웹XR ▲메터바스 비전 ▲비주얼 컴먼센스 리즈닝 ▲퀀텀머신러닝 ▲딥페이크 추적 ▲스마트 컨트랙트 렌즈 ▲리얼타임 머신러닝 ▲코그너티드 디지털 트윈 ▲3D 멀티 오브젝트 트랙킹 ▲오토머티브 날리지 디스커버리 ▲뉴로심볼릭AI ▲알고리즘 워페어 ▲프로그래머블 월드(Programmable World) 등이다.

4 데이터 오염 공격이란 학습 데이터를 조작해 AI 모델의 정확도를 낮추거나 오류를 유발하는 유형의 공격을 말한다.

5 소프트웨어정책연구소, 디지털 기술전망, 2023

6 분석 데이터는 아카이브(arXiv) 프리프린트 논문 메타 데이터 집합으로 2007년부터 2023년 7월까지 전 학제 분야 논문 약 230만 건(총 4MB)을 대상으로 하고, 기술의 분류를 위해서 기술분류가 잘돼 있는 특허 코드 26만 개를 활용하여 클러스터 100개를 AI의 유사도 매칭 기법을 활용하여 도출. SW정책연 "미래디지털기술 30개 선정···첫 독자 툴 사용", ZDNet Korea(2023.11.28.)

2036년까지 유망한 위크(Weak) 기술 6개는 ▲마이아바타 ▲액티버티 비전 ▲불완전한 데이티를 사용한 프리액티브 모델 ▲휴먼 디지털 트윈 ▲유나이티드 러닝 프로세스 ▲셀프 라이팅 소프트웨어가 꼽혔다.

이러한 새로운 디지털 기술 트렌드에서 눈에 띄는 가장 큰 특징은 AI 패러다임의 지속과 심화이며, 특히 Tiny AI, Human Digital Twin으로 대변되는 AI의 개인화 추세와 Activity Vision, Spatial Computing, Hyper-Realistic Persona는 가상세계인 메타버스의 일상화를 촉발하는 트렌드로 이해된다. 이러한 기술전망 분석결과를 바탕으로 미래 사회의 특성 3가지 키워드를 도출해보면, 첫째, 인공지능 기술이 사회의 모든 측면에서 증강을 제공해 인간의 능력을 향상시키고 일상의 경험을 개선하는 증강사회(Augmentation Society)의 개념이 중요해질 것으로 보인다. 둘째, 멀티모달의 다채로운 형식(Modality)이 인간의 시각, 청각, 느낌과 상상 등을 상호 연결하며 새로운 경험과 수요를 창출하는 디지털 육감(Digital Six Sense), 즉, 디지털 공감각(Digital Synesthesia)의 세상이 도래할 수 있다는 것이다. 마지막으로 전기 사용비용과 디지털 인프라와 컴퓨팅 파워제한 등을 극복하는 대안으로 퀀텀컴퓨팅(Quantum Computing)이 부상하는 것으로 탐지됐으며, 다양한 연구의 시도들이 진행되고 있어, 상상보다 퀀텀의 시간이 생각보다 멀지 않았다는 예측도 가능하다.

이러한 AI의 개인화와 일상화, 디지털 공감각 기술의 확산은 혁신과 새로운 수요 등 긍정적 결과물과 동시에 새로운 사회적 불안과 안보위협의 부산물을 발생시킬 것으로 보인다. 기존의 디지털 기술이 촉발하는 위협이 보안의 기술적 요인에서 기인했다면, 상술한 디지털 신기술 출현으로 야기될 수 있는 새로운 위협은 종래에 경험하지 못했고, 보안의 관점으로는 접근과 해결이 쉽지 않은 측면, 즉 문화 및 사회체제와의 결합을 통해 발생하는 안보라서 이슈 발생의 복잡성과 은닉성 그리고 공수비대칭성의 수위가 상대적으로 더 높아질 것으로 보인다.

먼저 사회적 인식과 소통의 다양성을 위협하는 AI로써 생성형 AI의 개인화와 일상화 트렌드는 기록된 증거의 신뢰성에 대한 사회적 의문과 비용을 발생시킬 것으로 보인다. 가짜 뉴스와 전략적 선전의 위협 사례는 2016년 미국 대통령 선거에서도 뚜렷하게 나타났으며, '딥페이크' 기술은 사회적 신뢰와 민주주의의 기반이 되는 공정한 정보 전달을 위협하고 안정성을 저해할 위험성을 지니고 있다. 2023년 3월 유포된 트럼프 전 대통령이 경찰에 연행되는 가짜 사진과 젤렌스키 우크라이나 대통령의 가짜 동영상이 대표적이다. 2022년에는 전체 온라인 트래픽의 약 47%가 봇에 의해 생성됐다. 이 봇들은 '의제 설정'을 조작하고, 대중의 의견과 주목을 바꾸는 능력을 갖추고 있다. Woolley & Howard(2016)는 또한, 이러한 봇들이 정치적 논쟁과 대중 의견에 큰 영향을 미치고 있다고 지적한다. 트롤 팜을 통한 공격, 정보의 품질 저하, 익명성을 이용한 의견 조작 등은 정치적 토론의 질을 저하시킬 위험이 있다. 앞서 살펴본 AI의 자율성과 학습 특성이 사회적 번식의 과정을 거치면서 증폭된다면 변화하는 환경에 능동적으로 적응하며, 초기에 설정된 알고리즘의 기술적 명세를 넘어서, 실제 상황의 다양성에 따라 동작을 유연하게 변경할 수 있게 될 것이다. 한편 인텔은 페이크 탐지 기술 '페이크캐처'를 개발했으며, 챗GPT를 개발한 OpenAI는 특정 글을 입력하면 AI가 썼을 가능성이 얼마나 되는지 알려주는 생성형 AI를 감지하는 '클래시파이어'를 출시하는 등 각국 정부와 기업들은 딥페이크 기술 대책을 마련하고 있으나, 문제는 정교한 조작 기술에 비해 아직 이를 잡아내는 탐지

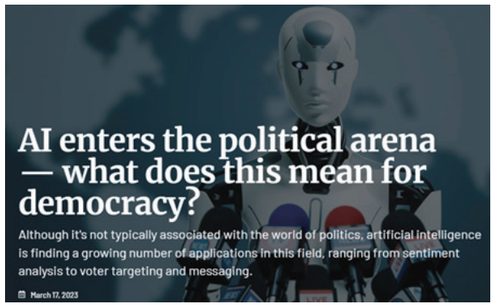
능력은 떨어진다는 점이다. 딥페이크 탐지 기술의 정확도는 아직 70%를 밑도는 수준이고, 조작 흔적을 찾을 수 있는 '디지털 포렌식' 전문가도 부족하다. 미국과 유럽 등에서는 AI를 이용해 만들어낸 콘텐츠에 'AI가 만들었다'는 식별 표시(워터마크) 부착과 AI의 윤리성의 강조를 추진하고 있지만, 불순한 목적을 가진 사람이나 세력은 얼마든지 이를 피할 수 있는 것이 현실이다.

둘째, AI가 대의민주주의 실패와 결합하면서 실체적 안보 위협으로 부상하는 새로운 안보 위협의 유형으로, 예컨대 참여와 피드백의 수월성을 극한의 수준으로 낮추는 디지털 기술에 의지하고 유권자의 의견과 사회적 다수의 요구라는 명분과 결합하면서 정치적 결정력을 갖는 새로운 형태의 권력으로 무장될 가능성이 바로 그것이다. 직접성(Directness), 직접접촉(Disintermediation), 상호성(Interactivity), 적응성(Adaptability), 즉각적 반응성(Instantaneous Responsiveness)의 디지털미디어와 플랫폼의 특성이 기존의 정치의 아날로그적 참여와 오프라인에서의 의견수렴과는 비교 불가의 가공할 효율성을 보이며 사회적 구성원들 간에 공감과 참여를 유발하며 진화한다면 새로운 AI 중심의 의사결정체제의 형성 가능성마저도 상상할 수 있을 것이다. 새롭게 등장해 주목을 받고 있는 디지털 정당과 의겸수렴체제로 미국 '체인지닷오아르지(change.org)'나 남미의 대표적 디지털 정다인 데모크라시OS를 예로 들 수 있다. 데모크라시OS는 아르헨티나의 사회운동가 피아 만치니와 산티아고 시리가 만든 SW로서 깃허브(Github)를 통해 소스코드를 공개하고 있으며, AI 챗봇을 당대표로 삼아 2022년 11월 덴마크 총선에 도전했던 소규모 정당 '신서틱 파티(Synthetic Party·인조정당)'에서는 '마이크로(Micro·초소형)' 사이즈의 정당이지만

머신러닝(Machine Learning) 모델을 통해 어떻게 사람들의 생각을 모으고 그걸 정치적인 비전으로 만들 수 있는지에 관한 가능성을 보여주는 것을 정당의 목표로 삼고 있다. 결과적으로 AI 챗봇이 당대표를 맡았다는 점과 챗봇 당대표가 내놓은 이 같은 파격적인 아이디어에도 불구하고 총선에서 의회에 입성하는 것에 실패했지만, 전 세계의 AI 정당과 가상정치인들과의 네트워크도 형성해나가는 중에 있다.

다양한 의견을 수렴하고 조합해 최상의 결론을 도출할 수 있다는 측면에서 디지털 정당체제는 장점을 가지고 있으나, 지배의 형태가 복잡해지고 의사결정의 과정 자체가 시스템 내부로 은닉되면서 의견과 요구의 실체적 진실 논란이 발생할 가능성이 있다. 또 이것이 걸러지지 않고 곧바로 사회적 의사결정을 연결될 가능성이 역시 존재한다. 따라서 향후 소수가 AI와 플랫폼의 독점성을 활용해 일반 시민을 알고리즘의 사용자에서 알고리즘의 희생자로 만드는 것에 대한 방어기제가 AI가 견인하는 미래 사회의 중요한 이슈가 될 것으로 전망된다.

[그림 3] AI 정치인의 등장 영상과 민주주의 체제에의 영향



4 요약과 시사점

이 글은 AI와 데이터 생태계의 기술특성을 살펴보고, 미래 디지털 기술의 전망을 기반으로 새롭게 부상하고 있는 안보의 새로운 유형 두 가지, 사회적 인식과 소통의 다양성을 위협하고 대의민주주의 실패와 결합하면서 실체적 안보 위협으로 부상하는 안보의 새로운 유형을 소개했다. 사실 디지털 기술 자체가 사회적 수요와 적용의 과정을 통해 끊임없이 진화하며 새로운 혁신을 창출하는 동시에 불안과 위협의 새로운 안보 이슈를 발생시키고 있어서, 앞으로의 디지털 안보 이슈는 기존의 기술적 접근과 더불어 사회적, 체제적 가치의 이슈와 맞물려 해법의 모색이 다소 복잡해지는 양상으로 치달을 확률이 높다. 일례로 데이터보안의 중요한 이슈인 데이터 생성의 주체도 그간 실제 생물학적 인간 중심이었다면, 실존적 자신을 대신하는 가상의 Avatar도 데이터 생성의 또 다른 주체로 등장하고 있으며, 실체적 인격과 특성이 가상 공간 기술과 만나면서 복수의 Digital Persona로 분화하며 진화하는 상황에서 데이터 보안의 근간이 되는 'Identified', 'Pseudonymized', 'Anonymized'의 프레임도 일정 정도 변화가 불가피할 것으로 보인다.

사실 MZ세대의 인식 속에 전화번호, 카카오톡 ID, 인스타그램의 ID는 이미 등가가 아니며, 생물학적 나와 가상의 페르소나가 가지는 인격, 영혼, 감정을 구분하고 공존시키려는 디지털 멀티 정체성의 세상이 펼쳐지고 있는 것이다. 향후 등장할 더욱 다양한 생성형 AI와 가상기술의 부상이 사회적 번식과 변의 과정을 거치게 되면 조만간 데이터 익명화에 대한 확고한 믿음이 흐려지고, 현재의 단층적 비식별 정보의 보안 수준과 의무를 실제와 가상 혹은 그 이상의 차원으로 구분해 다층 구조의 보안 수준과 차등적 의무를 부과하는 방식으로 전환될 수도 있어 보인다.

⊙ 참고자료

Saviotti, P. P. (1996). Technological evolution, variety and the economy. Books.

Jacobides, M. G., Brusoni, S., & Candelon, F. (2021). The evolutionary dynamics of the artificial intelligence ecosystem. Strategy Science, 6(4), 412-435.

Woolley, S. C., & Howard, P. N. (2016). Political communication, computational propaganda, and autonomous agents: Introduction. International journal of Communication, 10.

지디넷(2023.11.28.), SW정책연 "미래디지털기술 30개 선정···첫 독자 툴 사용"