FOCUS 1

모든 산업을 알아야 하는 산업: AI 시대 금융의 변화와 정책 제언

· **서병기** | UNIST 경영과학부 교수 | bkseo@unist.ac.kr

세 장면, 하나의 질문

<1씬> 앤트로픽의 미토스 소동

지난 4월, 앤트로픽(Anthropic)이 개발한 새로운 인공지능 모델인 '미토스(Mythos)'가 발표되자마자 작지 않은 소동이 일어났다. 미토스는 아직 아무도 모르는 보안 취약점, 이른바 제로데이(zero-day)를 스스로 찾아내는 녀석이다. 사람 손과 자동화 검사를 20년 넘게 빠져나간 것들이 걸려 나왔다. 가장 오래된 것은 보안으로 이름난 OpenBSD의 27년 묵은 버그였다. 회사 발표에 따르면, 한 달 남짓의 테스트에서 찾아낸 것만 수천 개다. 그런데 정작 흥미로운 것은 발표 다음에 벌어진 일이다. 24시간이 채 지나기 전에 미 연준 의장과 재무장관이 재무부 청사로 대형 은행 CEO들을 긴급 소집했다. 국방부도 정보기관도 아니고, 왜 하필 은행이었을까? 그 자리에서 내려진 주문이 또 의미심장하다. 미토스를 막으라는 것이 아니라, 미토스를 직접 써서 각자의 약점부터 찾으라는 것이었다. 파장은 곧 한국으로도 전해졌다. 금융위원회는 지난 5월, 2013년 이후 십수 년간 금융보안의 대원칙으로 지켜져 왔던 망분리 규제를 완화하겠다고 밝혔다. "AI 공격은 AI로 방어"한다는 것이 당국의 표현이었다. 막아야 할 대상이 곧 막을 수단이 된 셈인데, 그렇다면 규제는 대체 무엇을 규제해야 하는 것일까?

<2씬> 레오폴드의 몰락

지난 7월의 뉴욕. 시추에이셔널 어웨어니스(Situational Awareness)라는 헤지펀드가 마진콜을 맞았다. 2024년 여름 혜성처럼 등장해 2억 달러 남짓을 2년 만에 450억 달러까지 불린, 스물다섯 살의 천재 레오폴드 아셴브레너(Leopold Aschenbrenner)가 운용하던 펀드였다. AGI(Artificial General Intelligence)가 머지않아 도래한다는 전망 하나로 데이터센터와 반도체, 전력 인프라를 네 배 레버리지까지 사들인 결과였다. 그리고 한 달 만에 운용자산의 3분의 2가 사라졌다. 상장주식은 통째로 경쟁 헤지펀드에 10% 가까이 싼 값에 넘어갔고, 그 안에는 SK하이닉스도 있었다. 그 주 국내 증시가 이렇다 할 이유 없이 급락했던 것이 사실은 이 때문이었다는 분석도 있다. 그렇다고 이 펀드의 전망이 틀렸다고 판명된 것도 아니다. 미래를 옳게 보고도 무너질 수 있다면, 무너진 것은 대체 무엇인가?

<3씬> 금융 특화 파운데이션 모델

얼마 전 산업 특화 파운데이션 모델을 개발하는 공모 사업에 금융 주제로 참여해 보려고 계획서를 준비한 적이 있다. 그런데 금융에 특화된 파운데이션 모델이란 대체 어떤 물건이어야 할까? 궁리를 거듭하다 보니 아무래도 증권사 리서치센터 정도는 되어야겠다는 데 생각이 미쳤다. 리서치센터가 어떤 곳인가? 반도체와 조선, 유통과 제약과 건설을 각각 맡은 애널리스트들이 거의 모든 산업 섹터를 나누어 커버하는 조직이다. 거기까지 생각이 닿는 순간 계획서를 쓰던 손이 멈췄다. 아니, 이건 결국 범용 모델이 아닌가.

생각해 보면 당연한 귀결이었다. 금융은 스스로 무엇을 생산하지 않는다. 다른 산업이 벌이는 일을 자금과 가격의 형태로 되비추는 것이 금융이 하는 일이다. 조선사에 대출을 내주려면 조선업을 알아야 하고, 정유사의 회사채를 평가하려면 정유업을 알아야 한다. 금융을 잘하려면 모든 산업을 알아야 한다는 결론은 처음부터 정해져 있던 셈이다. 본 기고의 제목은 여기서 나왔다.

위의 세 장면은 각각 시스템과 인프라, 투자와 자산 가격, 그리고 고객을 마주하는 서비스의 문제를 건드리고 있다. 세 영역에서 공통으로 드러나는 문제는 AI가 판단의 속도와 범위를 빠르게 넓히는 동안, 그 판단을 확인하고 책임지는 체계는 아직 그 속도를 따라가지 못하고 있다는 점이다. 이 글에서는 세 영역에서 지금 무엇이 벌어지고 있는지를 가장 앞서 있는 곳들의 사례로 살펴보고자 한다. 우리가 앞으로 마주할 문제 대부분은 이미 그곳에서 한 번씩 모습을 드러냈다.

[표 1] 금융 AI 적용 영역

적용 영역	주요 내용	핵심 리스크	대응 방향
시스템	보안 관제, 이상 거래 탐지, 결제·정산 자동화, 감독 업무(SupTech)	공급자 집중, 사이버 위협의 고도화, 인프라 설계 전제의 붕괴	제3자 리스크 관리, 결제 인프라 재설계, 표준 형성 과정 참여
투자	로보어드바이저, 알고리즘 트레이딩, AI 리서치·리포트 생성	모델 동조화에 따른 쏠림, 레버리지 확대, 생성 정보의 신뢰성	자동주문 식별, 레버리지 익스포저 관측, 산출물 검증
서비스	신용평가 고도화, 상담·판매 채널, 언더라이팅, 내부 업무 자동화	설명책임 공백, 편향, 책임 소재의 분산	설명·이의제기 절차, 사람의 개입 통제, 책임 소재의 정리

1. 시스템 - 방어의 도구가 곧 위협의 도구일 때

미토스 사태가 드러낸 것은 새로운 해킹 기법이 아니다. 공격과 방어 사이의 비대칭이 사라졌다는 사실이다. 금융보안은 그동안 시간차 위에 서 있었다. 취약점을 찾는 시간, 그것을 악용하는 시간, 그 사이에 패치를 배포하는 시간. 망분리도 결국 이 시간을 벌기 위한 장치였다. 그런데 취약점이 알려진 뒤 그것을 뚫는

코드가 나오기까지 걸리던 며칠에서 몇 주가 몇 시간으로 줄면, 패치를 배포할 시간차 자체가 사라진다. 방어하는 쪽이 같은 모델을 쓰지 않으면 애초에 경기가 성립하지 않는 것이다.

진짜 문제는 그다음이다. 방어를 위해 고성능 AI를 쓴다는 것은, 그 모델을 만드는 소수의 회사와 그것을 얹어 돌리는 소수의 클라우드 사업자에게 금융 시스템의 안전을 맡긴다는 뜻이기도 하다. 금융안정위원회(FSB)가 AI 관련 취약성 가운데 가장 앞에 놓은 것도 이 제3자 의존과 공급자 집중 문제였다. 종래의 금융 리스크는 대체로 금융회사 안에 있었다. 자본과 유동성이 충분한지를 보면 됐다. 그런데 AI가 매개하는 리스크는 금융회사 밖에, 그것도 금융감독의 관할 밖에 놓인다. 감독의 대상은 은행인데, 정작 그 은행의 판단을 좌우하는 것은 감독 대상도 아닌 어느 기술회사의 모델 업데이트인 상황이 오는 것이다.

한편, 더 근본적인 변화는 결제와 정산 인프라에서 이미 시작되고 있다. 오늘의 금융 인프라는 예외 없이 사람을 전제로 설계되어 있다. 영업시간이 있고, 건마다 본인 확인을 하고, 하루 단위로 모아 정산한다. 그런데 사람이 아니라 AI 에이전트가 거래의 주체가 되면 어떻게 될까? 기계는 밤낮을 가리지 않고, 건당 금액은 몇 원 수준이지만 빈도는 초 단위다. 사람이 매 건을 승인한다는 것은 애초에 성립하지 않는 이야기다.

그 답으로 등장한 것이 x402라는 결제 규약이다. 이름의 유래가 재미있다. 웹의 통신 규약인 HTTP에는 402번 'Payment Required(결제 필요)'라는 상태 코드가 있다. 웹 초창기 설계자들이 훗날 값을 치르는 일이 생길 것을 내다보고 '장래를 위해 예약'만 해둔 코드인데, 카드 결제의 최소 단가 탓에 소액 결제가 성립하지 않아 30년 가까이 이렇다 할 쓰임을 얻지 못했다. 2026년 7월, 리눅스 재단이 이 코드를 되살린 x402 재단을 출범시켰다. 회원사 명단이 볼만하다. 비자, 마스터카드, 아메리칸익스프레스, 스트라이프, 아마존웹서비스. 카드망과 스테이블코인 진영이 같은 테이블에 앉은 것이다.

규모도 이미 실험 단계를 넘어섰다. 비자의 스테이블코인 결제는 2026년 4월 기준 연 70억 달러 수준이고, 마스터카드는 같은 해 3월 스테이블코인 결제 플랫폼을 최대 18억 달러에 인수하기로 했다. 아마존은 자사 에이전트 서비스에 x402를 넣어 건당 약 200밀리초, 1센트 미만의 비용으로 결제가 끝나게 했다. 사람이 물건을 사는 결제가 아니다. 에이전트가 연산과 데이터와 API 호출의 값을, 카드로는 감당이 안 되는 빈도와 단가로 치르는 것이다.

여기서 눈여겨봐야 할 것은 규칙이 정해지는 자리다. 기계가 돈을 주고받는 방식의 표준이 어느 나라 감독 당국도 아니고 국제기구도 아닌, 카드망과 클라우드와 암호자산 사업자가 함께 앉은 민간 기구에서 만들어지고 있다. 한국 기업도 카카오페이를 비롯해 세 곳이 이름을 올리기는 했지만, 규약의 방향을 의결하는 프리미어 등급 17개 자리에는 포함되지 않았다. 표준은 한번 굳으면 되돌리기 어렵다. 아직 논의가 열려 있는 지금, 참여의 수준을 한 단계 올려둘 필요가 있다.

규칙의 문제에서 한 걸음 더 들어가면 분배의 문제와 만난다. AI가 만들어내는 가치는 데이터를 제공한 쪽, 모델을 만든 쪽, 업무에 적용한 쪽에 걸쳐 발생하는데, 현재의 정산 구조는 이 사슬을 반영하지 못한다. 기여도에 따라 토큰으로 가치를 배분하자는 토크노믹스도, 자동화에 과세하자는 로봇세도, 결국 같은 질문에 대한 다른 답안이다. 누가 만든 가치를 누가 가져가는가? 이 질문에 답하지 않은 채 속도만 높이면, 그 이익이 어디로 흘러갔는지는 한참 뒤에야 확인하게 될 것이다.

2. 투자 - 옳은 전망과 감당할 수 있는 전망

다시 7월의 헤지펀드로 돌아가 보자. 이 펀드의 몰락을 두고 AI 낙관론의 파산이라고들 하지만, 정확한 진단은 아니다. AGI가 언제 오는지에 대한 이 펀드의 판단은 지금도 틀렸다고 확인된 바 없다. 무너진 것은 전망이 아니라 그 전망을 담은 그릇이었다. 간단한 산수를 해보자. 네 배 레버리지로 산 자산은 가격이 25%만 빠져도 자기자본이 전부 사라진다. 전망이 옳더라도 실현되기까지의 경로에서 그런 조정을 한 번이라도 겪으면 결과는 같다. 옳은 전망과 감당할 수 있는 전망은 전혀 다른 것이다.

이 대목에서 조금 뜻밖의 사실이 있다. 복수의 대형 투자은행이 신용을 공여했지만, 이번에는 손실을 본 곳이 없었다. 증거금을 올려 받고, 포지션을 통째로 다른 헤지펀드에 넘기는 것으로 정리가 끝났다. 2021년 아케고스(Archegos) 사태 이후 프라임브로커들이 다듬어 온 위험관리가 제 몫을 한 셈이다. 아케고스는 TRS(Total Return Swap)를 써서 포지션 자체를 감췄지만, 이번 펀드의 상장주식 보유는 공시에 드러나 있었다.

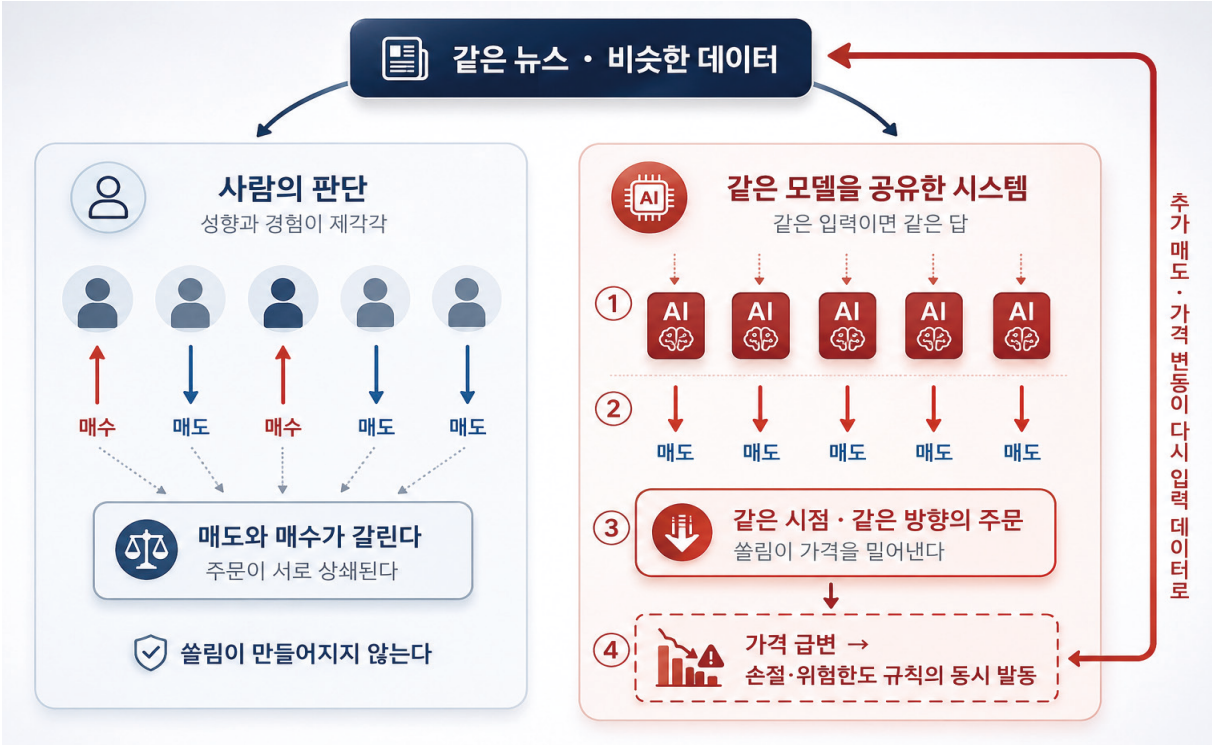
은행은 무사했는데 시장은 흔들렸다. 강제 매각이 진행되는 동안 국내 증시가 이유 없이 밀린 것이 그 증거다. 그리고 정리가 끝난 뒤에도 이 펀드에는 100억 달러 규모의 자산이 남았는데, 대부분이 값을 매기기 어려운 비상장 지분이다. 개별 금융회사의 장부가 안전해지는 것과 시장 전체가 안전해지는 것은 다른 문제다.

같은 시기, 투자은행 안에서 인공지능은 도구를 넘어 인력이 되어가고 있다. JP모건은 내부 생성형 AI 도구를 20만 명이 넘는 직원에게 배포했고, 이 회사의 최고 분석 책임자는 "장시간 자율 에이전트의 시대에 들어섰다"라며 에이전트가 한두 시간씩 스스로 일하고 앞으로는 며칠, 몇 주 단위로 갈 것이라고 했다.

골드만삭스는 자체 AI 어시스턴트를 4만 6천여 명 전 직원에게 열었고, 모건스탠리는 한 발 더 나가, 기업 고객의 AI 에이전트가 자사 주식 보상 관리 플랫폼에 직접 접속해 데이터를 가져가도록 여는 중이다. 사람 보라고 만든 화면을 거치지 않고 기계가 곧장 들어오는 구조다.

그런데 골드만삭스의 그 어시스턴트가 무엇으로 만들어졌는지를 들여다보면 이야기가 달라진다. 오픈AI와 구글, 앤트로픽의 모델 위에 골드만 자체의 보안 계층을 얹은 구조다. 세계에서 가장 앞선 투자은행조차 자기 모델을 갖고 있지 않은 것이다. 남들과 같은 소수의 파운데이션 모델을 쓰면서, 그 위에 자기 규율을 얹고 있을 뿐이다.

[그림 1] 모델 동조화에 따른 시장 쏠림



여기서 조용히 자라나는 문제가 모델 동조화다. 같은 모델에 비슷한 데이터를 넣으면 비슷한 답이 나온다. 평시에는 이것이 오히려 장점처럼 보인다. 판단의 품질이 고르게 유지되니까. 문제는 시장이 한쪽으로 기울기 시작할 때다. 사람의 판단이라면 성향과 경험이 제각각이어서 같은 뉴스에도 매도와 매수가 갈리지만, 같은 모델을 공유하는 시스템들은 같은 시점에 같은 방향으로 움직인다. 금융안정위원회가 시장 상관성과 쏠림의 증폭을 AI의 시스템 리스크로 지목한 것도 이 때문인데, 정작 감독 당국은 관측되는 쏠림 가운데 어디까지가 AI 탓인지 식별할 방법이 마땅치 않다고 토로한다.

이 문제는 우리에게 더 무겁게 다가온다. 골드만조차 남의 모델을 쓴다면, 자체 파운데이션 모델을 감당할 여력이 있는 국내 금융회사는 사실상 없다고 봐야 한다. 후발 시장일수록 같은 모델에 더 깊이 기대게 되고, 그만큼 판단이 닮아갈 여지도 크다. 동조화는 프런티어보다 오히려 우리 쪽에서 먼저 나타날지도 모른다.

여기에 하나가 더 겹친다. AI가 만들어낸 분석이 시장 가격에 반영되는 경로다. 종래의 애널리스트 리포트는 작성자의 이름과 소속, 평판이라는 담보를 달고 시장에 나왔다. 생성형 모델이 만들어낸 요약과 전망에는 그런 담보가 없다. 근거가 주장을 실제로 뒷받침하는지 확인할 길이 없는 채로 정보가 유통되고, 매매로 이어진다. 가격이란 결국 정보를 반영하는 장치인데, 그 정보의 출처를 검증할 수 없다면 가격이 반영하고 있는 것은 대체 무엇일까?

마지막으로 하나만 짚어두자. 요즘 투자라는 말은 두 방향을 동시에 가리킨다. AI로 투자하는 것과, AI에 투자하는 것. 그리고 후자의 규모는 이미 전자를 압도한다. 데이터센터와 반도체, 전력 설비로 흘러드는 자금은 상당 부분 차입에 의존하고 있으며, 국제결제은행(BIS)은 AI 관련 자산의 가치 평가가 실제 기업 이익보다 앞서 나가고 있다고 경고한 바 있다. 앞서 본 헤지펀드는 이 두 방향이 만나는 지점에서 무너졌다. AI를 활용해 AI에 베팅했는데, 그 베팅을 담아낼 그릇이 없었던 것이다.

투자 영역에서 지금 검토해 볼 만한 것은 레버리지와 자동 주문에 대한 통합적인 관측이다. 해외 펀드의 강제 청산이 국내 종목으로 옮겨붙는 경로가 어딘가에서는 관측되어야 하고, AI가 발주한 주문에는 그것을 식별할 표시가 있어야 한다. 그래야 시장이 밀릴 때 그것이 국내 사정 때문인지 밖에서 옮겨온 것인지, 그리고 그 가운데 AI에 기인한 부분이 얼마인지 사후적으로나마 분해해 볼 수 있다.

3. 서비스 - 특화될 수 없는 산업

고객을 마주하는 자리에서 인공지능이 가장 깊이 들어간 영역은 신용평가다. 종래의 신용평가는 금융거래 이력이라는 좁은 창으로 사람을 들여다보는 일이었다. 대출을 받아본 적도, 카드를 써본 적도 없는 사람은 위험한 사람이 아니라 알 수 없는 사람이다. 그런데 알 수 없다는 이유로 위험한 사람과 같은 취급을 받아왔다. 이른바 씬파일러(thin filer) 문제다.

미국의 온라인 대출 플랫폼 업스타트(Upstart)는 이 창을 넓히는 데 가장 멀리 간 사례로 꼽힌다. 회사 발표에 따르면 2,500개가 넘는 변수를 쓰고, 대출의 91%가 업스타트 쪽 사람 손을 거치지 않고 처리된다. 자사 모형을 가상의 전통적 심사 모형과 비교했더니 같은 손실률에서 44% 더 많은 신청자를 승인할 수 있었다는 것이 이 회사가 내세우는 수치다. 후불결제 회사 클라르나(Klarna)도 1억 명 이상의 고객을 상대로 심사를 1초도 안 되는 사이에 끝낸다. 보이지 않던 사람들이 보이기 시작한 것이다.

그런데 창이 넓어지면 설명해야 할 것도 함께 늘어난다. 2,500개의 변수를 비선형으로 결합한 모형을 붙들고 "이 사람이 왜 거절되었는가"를 물으면 어떤 답이 나올까? 정확도가 높아질수록 설명은 어려워진다는 오래된 긴장이 이제는 규제의 형태로 되돌아오고 있고, 각국의 대응은 갈린다. 미국은 신용 거절에 이유를 알릴 의무를 개별법에 두고, 감독 당국이 개별 사업자와 협의해 길을 터주는 방식을 병행해 왔다. 업스타트가 2017년 소비자금융보호국(CFPB)으로부터 받은 노액션 레터(No-Action Letter)가 그 예다. 반면 유럽은 신용평가를 고위험으로 분류해, 시장에 들어가기 전에 요건을 채우게 하는 쪽을 택했다.

[표 2] 신용평가에 대한 각국의 접근

구분	미국	EU	한국
근거	ECOA·FCRA 등 개별법의 거절 사유 고지 의무	AI Act, GDPR 제22조 (자동화된 결정)	인공지능 기본법 및 시행령
신용평가의 지위	별도 분류 없이 기존 소비자보호법제 적용	고위험(high-risk)으로 분류	고영향 영역으로 분류
규제 방식	사후 설명 중심, 감독 당국과의 개별 협의(노액션 레터) 병행	사전 규율, 사람의 개입과 절차적 권리 요구	사전 규율에 가까움, 위험·영향평가와 문서화 의무
대표 사례	업스타트(CFPB 노액션 레터, 2017년 발급·2022년 종료)	레볼루트 (리투아니아 중앙은행 감독 리뷰)	시행 초기 단계

유럽의 방식이 실무에서 어떤 모습이 되는지는 레볼루트(Revolut) 사례가 잘 보여준다. 리투아니아 중앙은행은 이 회사의 신용 모형을 검토하면서 설명가능성과 사람의 개입 통제를 강화하라고 요구했고, 회사는 거절 사유를 변수 기여도로 환산해 제시하고 일정 기준을 넘는 거절 건은 사람이 다시 들여다보는 대기열을 두는 것으로 답했다. 규제가 문서 한 장으로 끝나지 않고 업무 절차의 재설계로 이어진 것이다. 우리의 인공지능 기본법도 대출 심사를 고영향 영역으로 분류해 두었으니, 국내 금융회사들도 머지않아 같은 종류의 재설계를 요구받게 된다.

보험 쪽에서는 성격이 다른 문제가 자라고 있다. 인공지능이 언더라이팅과 보험금 심사를 돕는 것은 이제 새로운 이야기가 아니다. 정작 어려운 것은 AI 자체를 담보로 삼는 상품이다. 자율주행의 판단 오류, 의료 영상 판독의 실수, 투자 자문 알고리즘의 손실. 누가 어디까지 책임지는가? 위험을 인수하려면 그것이 얼마나 자주, 어느 크기로 발생하는지 알아야 하는데, 그럴 데이터가 아직 세상에 없다. 손해율을 모르는 위험에는 값을 매길 수 없고, 값을 매길 수 없는 위험은 거래되지 않는다.

앞서 모건스탠리가 외부 에이전트에 플랫폼을 열고 있다고 했는데, 사실 이것이 서비스 영역에서 일어날 가장 큰 변화의 예고편이다. 고객이 자신의 AI 비서에게 "조건 괜찮은 대출 좀 알아봐 줘"라고 말하고, 그 비서가 여러 금융회사를 비교해 신청까지 마치는 구조가 자리 잡으면 어떻게 될까? 금융회사가 마주하는

상대는 더 이상 사람이 아니라 다른 기계다. 상품 설명 의무는 누구에게 이행해야 하는가? 적합성 원칙은 누구를 기준으로 판단하는가?

이 대목에서 미리 정리해 둘 것이 책임의 소재다. 모델을 만든 회사, 그것을 업무에 쓴 금융회사, 고객을 대신해 거래를 실행하는 에이전트 사업자. 이 셋 사이에서 손해가 났을 때 누가 어디까지 부담하는지 정해지지 않으면, 보험은 값을 매기지 못하고 금융회사는 도입을 주저하게 된다.

돌고 돌아 이야기는 다시 한곳으로 모인다. 조선사의 대출을 심사하려면 수주 잔고를, 정유사의 회사채를 평가하려면 정제 마진을, 한 사람의 신용을 보려면 그가 속한 업종과 지역의 사정까지 읽어야 한다. 금융의 서비스는 예외 없이, 다른 산업과 다른 삶에 대한 이해 위에서만 성립한다.

그렇다면 금융회사의 경쟁력은 어디에서 나오는가? 적어도 모델 자체에서 나오지는 않는다는 것이 필자의 잠정적인 답이다. 골드만이 그랬듯, 남들과 같은 모델을 쓰면서도 그 산출물을 검증하고 관리하는 능력에서 나온다. 어떤 데이터를 붙였는가, 그 판단을 어떻게 확인하는가, 틀렸을 때 무엇이 작동하는가. 특화는 모델에 있지 않고, 그 위에 얹는 검증 체계에 있다.

도입의 속도, 검증의 능력

세 영역을 지나오며 마주친 문제들은 결국 하나로 모인다. 판단은 빨라지는데, 그 판단을 확인할 방법이 따라오지 못하고 있다는 것이다. 영역마다 하나씩 짚어둔 것 외에, 두 가지만 덧붙이고 싶다.

하나는 공통의 기준이다. AI가 생성한 금융정보를 무엇으로 믿을 만하다고 판단할 것인가? 지금의 논의는 대체로 모델의 성능과 개발 과정을 향해 있다. 그러나 시장에 실제로 영향을 미치는 것은 모델이 아니라 그 결과로 나온 산출물이다. 근거를 추적할 수 있는가? 그 근거가 주장을 실제로 뒷받침하는가? 사실과 다른 내용이 어느 정도 섞여 있는가? 이런 항목에 공통의 척도와 제3자 검증 절차가 마련된다면, 리포트든 상품 설명이든 비로소 신뢰의 근거를 갖게 될 것이다.

다른 하나는 검증을 위한 공동의 인프라다. 개별 회사가 각자 검증 체계를 갖추는 것은 대형사에게도 부담이고, 중소형사에게는 사실상 불가능한 일이다. 대학과 금융회사, 감독기관이 함께 쓰는 시험 환경을 두고, 실제 데이터를 대신할 합성 금융 데이터의 활용 기준까지 함께 정리한다면, 개인정보 제약과 데이터 부족이라는 두 병목을 한 번에 다뤄볼 수 있다. 이때 감독기관은 최소한의 공통 기준과 시험 환경을 마련하고, 실제 검증은 금융회사와 대학·전문기관 등 독립된 주체가 함께 수행하는 구조가 현실적일 것이다. 모든 AI 활용을 일률적으로 규율하기보다 소비자나 시장에 미치는 영향이 큰 영역부터 검증 수준을 높여가는 방식도 생각해 볼 수 있다.

오해를 피하기 위해 덧붙이자면, 위의 이야기들은 규제를 하나 더 얹자는 것이 아니다. 오히려 그 반대에 가깝다. 확인할 수 있는 것이 늘어날수록 풀어줄 수 있는 여지도 함께 넓어지기 때문이다. 무엇이 일어나고 있는지 모르는 상태에서는 규제를 강화하자는 쪽도 완화하자는 쪽도 근거를 대기 어렵다.

다시 첫 장면으로 돌아가 보자. 십수 년을 버텨온 규제가 모델 하나에 흔들렸을 때, 당국이 내린 결론은 그 모델을 막는 것이 아니라 쓰는 것이었다. 원칙의 후퇴라고 볼 수도 있겠지만, 필자는 오히려 정직한 판단이었다고 본다. 막을 수 없는 것을 막겠다고 버티기보다는, 쓰되 확인할 수 있는 체계를 만드는 편이 낫다. 문제는 쓰기로 한 속도만큼 확인하는 능력이 함께 자라고 있는가에 있다.

금융은 특화될 수 없는 산업이다. 다른 산업이 벌이는 모든 일이 자금과 가격이라는 형태로 이 산업에 모여들기 때문이다. 그래서 금융에 인공지능을 들이는 일은 신기술 하나를 도입하는 일이 아니라, 세상 전체를 읽어내는 장치를 사회의 자금 배분 기구에 얹는 일에 가깝다. 그 장치가 무엇을 근거로 판단했는지 물을 수 없다면, 우리는 이유도 모르는 채로 자금이 흘러가는 사회를 갖게 된다.

경쟁은 이미 도입의 속도에서 검증의 능력으로 옮겨가고 있다. 남들보다 먼저 쓰는 회사가 아니라, 남들과 같은 것을 쓰면서도 그 결과를 확인할 수 있는 회사가 결국 앞선다. 이번 국면이 우리 금융이 그 능력을 키워가는 계기가 되기를 바란다.

참고문헌

[1] Anthropic Frontier Red Team, “Assessing Claude Mythos Preview’s cybersecurity capabilities,” 2026. 4. 7. https://www.anthropic.com/research/mythos-preview

[2] AI Security Institute(UK), “Our evaluation of Claude Mythos Preview’s cyber capabilities,” 2026.

[3] Centre for Emerging Technology and Security(CETaS), “Claude Mythos: What Does Anthropic’s New Model Mean for the Future of Cybersecurity?,” 2026.

[4] 금융위원회, 망분리 규제 완화 및 금융보안 대응 관련 보도자료, 2026. 5.

[5] 미 재무장관·연준 의장의 대형은행 CEO 긴급 소집 및 Project Glasswing 관련 보도(CNBC, Bloomberg 등), 2026. 4.

[6] 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 및 동법 시행령, 2026. 1. 시행.

[7] Financial Stability Board, 인공지능과 금융안정 관련 보고서, 2024; 책임 있는 AI 도입을 위한 모범관행 협의보고서, 2026. 6.

[8] Bank for International Settlements, Annual Economic Report 중 인공지능 관련 장, 2026.

[9] Linux Foundation, x402 재단 출범 관련 발표 및 관련 보도, 2026. 7.

[10] Visa·Mastercard·AWS의 스테이블코인 결제 및 에이전트 결제 관련 발표와 보도, 2026.

[11] JPMorgan Chase의 AI 에이전트 도입 관련 보도(CNBC 등), 2026. 6.

[12] Goldman Sachs GS AI Assistant 관련 보도, 2025~2026; CNBC, Morgan Stanley의 외부 AI 에이전트 접속 개방 보도, 2026. 6. 3.

[13] Upstart 및 Klarna의 신용평가 모형 관련 공시·발표 자료; CFPB No-Action Letter(2017).

[14] Bank of Lithuania의 Revolut 신용모형 감독 검토 관련 자료, 2024.

[15] 시추에이셔널 어웨어니스 마진콜 관련 보도, 2026. 7.