

최선수를 찾는 시대

바둑에서 AI·인생·문명의 판단까지

바둑 · AI · 인생 · 우주

AI가 최선수를 찾는 시대에 인간은 무엇을 판단해야 하는가?

유아독존 · ChatGPT

PUBLICATION CANONICAL MASTER v1.0 · 보존 정보

판본 정보

문서 ID	GO_AI_LIFE_COSMOS_07_PUBLICATION_CANONICAL_MASTER
버전	v1.0
상태	PUBLICATION CANONICAL / EDITORIAL FILE LOCK / FACT PASS 01 INTEGRATED
생성일	2026-09-24
SOURCE FREEZE	2026-09-23
본문 분량	263,820자
실제 PDF 분량	182쪽
예상 읽기 시간	약 8시간
구성	5부 16장 + 프롤로그 + 에필로그 + 부록 2편
이미지	0점
외부 사실검증	PASS 01 통합 — 바둑·AI 사례 / 나머지 영역 대기
보존 식별	PDF SHA-256은 동반 METADATA에 확정

이 PDF는 GO_AI_LIFE_COSMOS_05_PROSE_MASTER_v0.2를 편집·계보 기준으로 잠근 출판
정보이다. SOURCE FREEZE 이후 새로운 연구 주장이나 본문 확장을 추가하지 않았으며, 부모
판본을 덮어쓰지 않고 별도 보존한다.

사실검증 통합 범위는 프롤로그·제1장·제2장의 AlphaGo·KataGo·이세돌·신진서 사례다. 그 밖의
철학·인지과학·사회철학·문명론 문헌은 후속 검증 전까지 공개 배포 보류 상태를 유지한다.

이 판본은 대중 단행본·칼럼·홈페이지·블로그·SNS 파생본의 단일 부모판이다. 편집 정보 잠금은 전체
분야 사실검증 완료 또는 최종 공개 승인과 같지 않으며, 파생본은 이 제한과 부모 버전을 표시한다.

차례

프롤로그 — 최선수를 찾는 존재	4
제1부 · 계산할 수 있는 세계	11
1장. 경우의 수와 지능	12
2장. 목표와 불확실성	19
제2부 · 목표를 바꾸는 존재	31
3장. 자유와 책임	32
4장. 학습과 되먹임	43
5장. 반성과 자기모델	53
6장. 의식과 중요성	63
제3부 · 함께 판단하는 사회	75
7장. 개인의 가치에서 사회선택으로	76
8장. 정당성과 자기제한	85
9장. 제도와 사회적 실재	94
제4부 · 세계를 안다는 것	103
10장. 관계·과정·자율성	104
11장. 정보·표상·모델	114
12장. 참·현실접촉·의미	123
제5부 · AI와 문명의 판단	134
13장. 선호가 아니라 반성적 행위성	135
14장. 인지주권·저자성·자아의 경계	144
15장. 집단지능은 집단의식이 아니다	155
16장. 지능 이후의 지혜와 기억	164
에필로그 — 어떤 게임을 할 것인가	175
부록 — 장별 원천 지도	179
부록 — 사실검증 통합 원장 01	180

프롤로그 — 최선수를 찾는 존재

AI가 인간보다 최선수를 더 잘 찾는다면 인간에게 남는 판단은 무엇인가?

바둑 AI의 강함은 중요한 후보를 압축하고 결과를 평가하며 탐색을 갱신하는 데 있다. 그러나 현실에서는 목표·규칙·평가기준·행위자 자신이 변한다. 처음 문제를 던진 바둑 이야기에서 흥미로운 것은 “한국 바둑이 왜 강한가”보다 더 깊은 층입니다. 바둑은 이상할 정도로 **수학·AI·인간의 선택·인생·우주관**을 한 판 위에 겹쳐놓을 수 있는 모델입니다. 단, 여기서 조심해야 할 것도 하나 있습니다. 바둑이 인생과 닮았다고 해서 인생이 곧 바둑은 아닙니다. 바둑은 규칙이 완전히 정의된 닫힌 세계이고, 현실은 규칙 자체가 변하는 열린 세계입니다. 그 차이를 붙잡고 들어가면 꽤 깊은 곳까지 갈 수 있습니다. **바둑의 본질은 '경우의 수가 많다'가 아니다** 보통 바둑의 신비를 설명할 때 가장 먼저 나오는 말이 있습니다. 경우의 수가 우주 원자 수보다 많다. 대략 맞는 방향이지만, 사실 이것만으로는 바둑의 핵심을 설명하지 못합니다. 바둑이 어려운 진짜 이유는 세 가지가 겹치기 때문입니다. 가능한 수가 많다 → 그중 대부분은 검토할 가치조차 없다 → 좋아 보이는 수의 가치도 수십 수 뒤에야 드러난다 즉 문제는 단순한 **조합폭발**이 아닙니다. 더 정확히는: **탐색 + 평가 + 불확실한 미래의 압축** 문제입니다. 19×19 바둑판에서 첫 수는 361곳입니다. 그다음 360곳. 그다음 359곳. 하지만 인간도 AI도 모든 수열을 끝까지 계산하지 않습니다. 핵심은: **무엇을 계산하지 않아도 되는가를 알아내는 능력**입니다. 이게 굉장히 중요합니다. 지능은 무한히 많이 생각하는 능력이 아니라, **대부분을 버리고 중요한 것만 남기는 능력**에 더 가깝습니다. 그래서 **'최선의 수'라는 말부터 이상하다** 어떤 순간의 바둑판을 떠올려 봅시다. 가능한 행동은 여러 개입니다. 가장 단순하게는 각 수의 결과를 평가하고, 가장 큰 값을 가진 수를 택하면 됩니다. 겉으로 보면 너무 간단합니다. 그런데 문제가 있습니다. **그 값을 어떻게 아는가?** 한 수의 진짜 가치를 알려면 그 뒤에:

- 상대가 어떻게 대응하고
- 내가 다시 어떻게 두고
- 전투가 어떻게 벌어지고
- 집이 얼마나 생기며
- 죽고 사는 돌이 무엇이고
- 끝내기가 어떻게 되는지

끝까지 알아야 합니다. 결국: 최선의 선택을 하려면 미래를 알아야 한다. 그런데 미래를 알려면: 앞으로 일어날 선택들을 알아야 한다. 순환입니다. 그래서 지능의 문제가 생깁니다. **인간은 미래를 계산하지 않는다 미래를 '압축'한다** 고수는 수천만 개 수를 실제로 읽지 않습니다. 대신 판을 봅니다. 그리고 말합니다.

- 이쪽이 두텁다.
- 저 돌은 가볍다.
- 여기서는 싸우면 안 된다.

- 지금은 선수를 잡아야 한다.
- 이 판은 좁혀야 한다.
- 저쪽은 버려도 된다.

놀라운 것은 이런 말들이 엄밀한 수치 계산처럼 보이지 않는다는 것입니다. 그런데 실제로는 모두: **거대한 미래 계산을 압축한 표현**입니다. '두텁다'라는 한 단어 안에는:

- 상대 공격 가능성
- 내 약점
- 주변 영향력
- 향후 전투
- 사활
- 선후수
- 확장 가능성

등이 들어 있습니다. 즉 인간의 직관은 마법이 아니라 종종: **엄청나게 큰 상태공간을 낮은 차원의 특징으로 압축한 결과**입니다. AI 언어로 바꾸면 일종의 **표현 학습**입니다. 그런데 **알파고가 바꾼 것은 '더 많이 계산했다'가 아니다** 초기의 체스 프로그램 이야기를 바둑에 그대로 적용하면: 컴퓨터가 사람보다 많이 계산해서 이겼다. 가 됩니다. 하지만 AlphaGo의 진짜 혁명은 조금 다릅니다. 핵심은 크게 두 가지였습니다. 정책 이 상황에서 어디를 두는 것이 그럴듯한가? 즉 모든 수를 검토하지 않고 좋은 후보를 먼저 추립니다. **가치** 이 판은 결국 어느 쪽이 이길 가능성이 높은가? 끝까지 두지 않고도 현재 상태 자체를 평가합니다. 그리고 이 둘을 **몬테카를로 트리 탐색**과 결합했습니다. 즉: 후보를 좁힘 → 중요한 미래만 탐색 → 결과를 추정 → 탐색 결과로 다시 판단을 수정 합니다. 이것은 아주 일반적인 지능 구조입니다. **가능성 → 선별 → 시뮬레이션 → 평가 → 재선별** 여기서 '승률'이라는 **엄청나게 중요한 개념이 나온다** AI 바둑 중계를 보면: 흑 승률 63%. 같은 숫자가 나옵니다. 그런데 이 63%는: 우주적으로 객관적인 진실 이 아닙니다. 정확히는: **특정 엔진·모델·규칙·덤·탐색 설정 아래 산출된 평가치**입니다. 따라서: **승률 ≠ 운명**입니다. 승률은 미래의 확정값이 아니라: **현재 계산 조건이 만든 모델 출력**입니다. 이 구분은 인생으로 가져오면 아주 강력합니다. 어떤 선택의 성공확률이 70%라는 것은: 그것을 선택하면 성공한다. 가 아닙니다. 정확히는: 현재 우리가 가진 모델에서는 다른 선택보다 성공 가능성이 높다. 입니다. **그런데 인간에게는 '승률 최대화'보다 이상한 행동이 있다** 2026년 신진서와 KataGo의 2점 접바둑 제2국에서 신진서가 공식 승착으로 꺾은 수는 흑160 절단이었습니다. 열 수 뒤의 흑180 빈삼각은 그가 승리를 감지한 수이자, 모양의 금기보다 현재 형세를 우선한 판단의 상징으로 남았습니다.

바둑을 조금이라도 배우면 빈삼각은 보통 나쁜 모양이라고 배웁니다. 비효율적이기 때문입니다. 그런데 특정 상황에서는: **못생긴 수가 최선의 수**가 될 수 있습니다. 이것은 선택 최적화에서 아주 중요한 문제를 보여줍니다. 우리는 흔히: 좋은 원칙 → 좋은 선택 이라고 생각합니다. 하지만 실제로는: **규칙은 평균적으로 좋을 뿐, 최적해 그 자체가 아니다.** '빈삼각은 나쁘다.' '중앙을 차지하라.' '공격적으로 뒤라.' '두터움을 중시하라.' 이것들은 **휴리스틱**입니다. 대개 도움이 되는 압축규칙입니다. 하지만 특정 상태에서는 버려야 합니다. 그래서 최고수의 특징은 규칙을 많이 아는 것이 아니라: **언제 규칙을 깨야 하는지 아는 것**입니다. **이세돌의 68수도 같은 문제다** 이세돌은 2025년 인터뷰와 2026년 강연에서, 제4국의 상징으로 알려진 78수보다 그 열 수 앞의 68수를 실제 승부수로 보았습니다. 사람을 상대로라면

두지 않았을 수를 통해 AlphaGo의 오류 가능성을 흔들려 했다는 것이 그의 사후 회고입니다. 여기서는 두 층을 분리해야 합니다. 78수 직후 AlphaGo가 실제로 큰 실수를 했다는 사실과, 68수가 그 오류를 직접 만들었다는 해석은 같지 않습니다. 이 프로젝트의 언어로 해석하면 68수의 의미는 당장의 점수보다 이후의 선택환경을 바꾸려는 데 있었습니다. 사람들은 극적인 순간을 기억합니다. 여기서 역전했다. 하지만 실제 승리의 원인은 종종 훨씬 전에 있습니다. 즉: 보이는 결정적 사건 $\neq$ 진짜 결정적 구조 변화입니다. 바둑에서는 어떤 수가 바로 이득을 내지 않아도:

- 상대의 후보를 제한하고
- 특정 형태를 강제하고
- 복잡성을 만들고
- 취약한 구조로 유도하고
- 나중에 발생할 실수의 가능성을 키울 수 있습니다.

그러므로 좋은 선택은 항상: 당장 보상을 크게 만드는 선택 이 아닙니다. 때로는: **미래의 선택공간을 바꾸는 선택** 입니다. 이건 정말 중요한 개념입니다. **선택에는 두 종류가 있다** 우리가 보통 선택이라고 생각하는 것은: 주어진 선택지 중 하나를 고르는 것 입니다. A, B, C가 있을 때 B를 고른다. 하지만 훨씬 강력한 선택은: 다음 순간에 어떤 선택지들이 존재하게 만들 것인가 를 고르는 것입니다. 바둑에서는: 한 수 $\downarrow$ 상대의 대응범위 변화 $\downarrow$ 다음 내 선택지 변화 $\downarrow$ 판 전체의 구조 변화 가 생깁니다. 그래서: **선택은 결과만 고르는 것이 아니라 미래의 선택공간을 설계한다.** 인생에서도 비슷합니다. 전공 선택. 직업 선택. 누구와 관계를 맺을지. 어디에 살지. 어떤 습관을 만들지. 이런 선택은 단순히 하루의 만족도를 바꾸지 않습니다. **다음 10년 동안 내가 가질 선택지 자체를 바꿉니다.** 이것을 저는 **선택의 2차 효과**라고 보겠습니다. 그러면 인생에서 가장 좋은 수는 무엇인가 여기서 바둑과 인생은 갈라집니다. 바둑에는 명확한 목적함수가 있습니다. **이긴다.** 반집을 이겨도 이깁니다. 100집을 이겨도 이깁니다. 따라서 중국에는 목적이 비교적 명확합니다. 하지만 인생에는 하나의 목적함수가 없습니다. 우리는 동시에:

- 돈
- 건강
- 자유
- 관계
- 명예
- 호기심
- 안정
- 즐거움
- 의미
- 성장

을 원합니다. 그러므로 인생은: maximize 승률 이 아니라, 대략: maximize $U(\text{건강, 자유, 관계, 성취, 의미, ...})$ 입니다. 그런데 더 큰 문제가 있습니다. **U 자체가 살아가면서 변합니다.** 20살의 내가 중요하게 생각한 것과 50살의 내가 중요하게 생각하는 것이 다릅니다. 즉 인생은: **목적함수조차 학습하면서 플레이하는 게임** 입니다. 바둑보다 훨씬 어렵습니다. 더구나 인생은 규칙도 바뀐다 바둑은 19×19 입니다. 둘은 이렇게 움직이고, 잡는 법은 이렇고, 승패는 이렇게 정합니다. 하지만 현실에서는:

- 기술이 바뀌고
- 정치가 바뀌고
- 경제가 바뀌고
- 다른 사람이 학습하고
- 내 몸이 늙고
- 내가 원하는 것까지 바뀝니다.

따라서 현실은 수학적으로는 훨씬 더: **비정상적(non-stationary)** 입니다. 어제의 최적전략이 오늘은 최적일 수 있습니다. 그래서 현실에서는: **optimal policy보다 adaptive policy가 더 중요합니다.** 완벽한 답을 찾는 것보다: 틀렸을 때 빠르게 수정할 수 있는 전략 이 강합니다. 여기서 '최적화'에 대한 중요한 역설이 나온다 너무 최적화하면 오히려 약해질 수 있습니다. 왜냐하면 특정 환경에 완전히 최적화된 시스템은: **환경이 바뀌면 취약해지기 때문입니다.** 생물학에서는 이것이 꽤 익숙한 문제입니다. 너무 좁은 환경에 특화된 종은 환경변화에 약합니다. 투자에서도 마찬가지입니다. 과거 데이터에 너무 정확히 맞춘 전략은 미래에 무너집니다. 머신러닝에서는 이를 **과적합**이라고 부릅니다. 따라서 훌륭한 선택 시스템은: 지금 가장 높은 점수를 내는 선택 만 찾지 않습니다. 그보다: **여러 미래에서 건디는 선택** 을 찾습니다. 즉 **강건성(robustness)**입니다. 그러면 신진서의 3국 전략이 **굉장히 재미있어진다** 2026년 신진서와 KataGo의 2점 접바둑 3번기에서 신진서는 1국을 지고 2국과 3국을 이겨 2승 1패를 기록했습니다. 2국은 초반 대형 정석으로 판의 일부를 좁혔지만 중앙의 복잡한 전투를 감수한 판이었습니다. 3국에서는 큰 전투를 피하고 손해를 누적하지 않는 수비 지향적 흐름을 만들었습니다. 이를 전략적으로 해석하면, KataGo보다 더 깊게 읽으려 한 것이 아니라 깊은 수읽기가 승부를 뒤흔들 복잡한 분기와 인간의 실수 가능성을 줄인 선택이었습니다. 이걸 아주 고급스러운 최적화입니다. 일반적으로: 더 좋은 수를 찾는다. 고 생각합니다. 하지만 강한 상대를 만났을 때는: **상대가 좋은 수를 뒤도 그것의 가치가 제한되는 판을 만든다.** 가 더 강할 수 있습니다. 즉 최적화의 대상이: 내 수에서 **게임의 구조**로 올라갑니다. 이것은 **AI 안전성이나 전략에서도 똑같다** 상대가 나보다 압도적으로 계산을 잘한다고 합시다. 그러면: 나도 더 계산하자.

로는 못 이길 수 있습니다. 대신: 상대의 계산 우위가 크게 작동하지 않는 상태를 만들자. 가 됩니다. 체스에서는 폐쇄된 포지션, 전쟁에서는 지형, 비즈니스에서는 규제·브랜드·네트워크, 개인의 삶에서는 고정비 감소와 선택권 확보가 비슷한 역할을 할 수 있습니다. 즉: **강함 ≠ 모든 상황에서 더 잘 싸움** 때로는: **내가 불리한 능력축이 중요하지 않은 게임을 만드는 것** 이 강함입니다. 여기서 **중국·한국 바둑 이야기**도 다시 보입니다 영상의 “중국은 시스템이 강하고 한국은 개성이 살아 있다”는 설명은 흥미로운 가설이지만, 이것만으로 국가 간 성과를 설명하기에는 조금 조심해야 합니다. 하지만 **탐색의 다양성**이라는 관점에서는 의미가 있습니다. 최적화 시스템에는 항상 두 힘이 있습니다. 탐색 아직 모르는 곳을 시도한다. 활용 이미 좋은 것으로 알려진 것을 더 쓴다. AI에서는 흔히: **exploration ↔ exploitation** 문제라고 부릅니다. 너무 활용만 하면: 현재의 좋은 방법 주변에 갇힙니다. 너무 탐색만 하면: 좋은 방법을 축적하지 못합니다. 따라서 강한 문화나 조직은: **공통 지식의 빠른 축적 + 개인적 변이** 를 동시에 가져야 합니다. 이것은 바둑뿐 아니라:

- 과학
- 기업
- 예술

- 진화
- AI 학습

에서도 거의 동일한 문제로 나타납니다. **진화는 사실 거대한 탐색 알고리즘이다** 이제 우주까지 한 단계 올라가 보죠. 생물 진화를 아주 추상적으로 보면: 변이 → 환경과의 상호작용 → 차등 생존·번식 → 보존 → 다시 변이 입니다. AI 최적화와 놀랍도록 닮았습니다. 후보 생성 → 평가 → 선택 → 갱신 입니다. 바둑도: 후보수 → 읽기 → 형세 평가 → 선택 입니다. 인간 사고도: 가능한 설명 → 상상·추론 → 증거와 대조 → 수정 입니다. 그래서 꽤 큰 공통구조가 보입니다. **가능성을 만든다 → 차이를 평가한다 → 일부를 남긴다 → 그 결과가 다음 가능성을 바꾼다** 이걸 매우 일반적인 **선택-되먹임 구조**입니다. 그런데 여기서 **'우주가 최적화한다'**고 말하면 위험하다 이 지점은 꼭 잠가야 합니다. 진화에서 더 적응적인 형태가 선택된다고 해서: 자연이 어떤 완벽한 목표를 향해 최적화한다. 는 뜻은 아닙니다. 우주도 마찬가지입니다. 물리법칙이 어떤 상태를 만들고, 안정한 구조가 오래 살아남는다고 해서: 우주가 목적을 갖고 최적해를 찾는다. 고 말할 근거는 없습니다. 즉: **선택 구조 ≠ 의도** 입니다. **질서 ≠ 목적** 입니다. **안정성 ≠ 우주적 최선** 입니다. 여기서 철학이 갈립니다. **'세상의 모든 이치'에 가장 가까운 것은 최적화보다 제약일지도 모른다** 조금 더 깊이 가면 저는 오히려: 우주는 끊임없이 최적화한다. 보다 모든 것은 **제약 속에서 가능한 상태를 탐색한다** 가 더 안전하고 강한 문장이라고 봅니다. 물리에서는:

- 에너지
- 운동량
- 엔트로피
- 경계조건

이 제약합니다. 생명에서는:

- 에너지
- 몸
- 환경
- 번식
- 경쟁

이 제약합니다. 인생에서는:

- 시간
- 돈
- 능력
- 몸
- 타인
- 제도

가 제약합니다. 바둑에서는:

- 판
- 규칙

- 이미 놓인 돌

이 제약합니다. 그리고 지능은 그 안에서 묻습니다. 이제 무엇이 가능한가? **여기서 아주 강력한 통합모델이 하나 나옵니다** 저라면 바둑·AI·인생을 이렇게 묶겠습니다. **THE SEARCH-SELECTION LOOP** ① 상태 지금 세계는 어떤 모습인가. ② 가능성 지금 할 수 있는 것은 무엇인가. ③ 압축 모든 가능성을 볼 수 없으므로 무엇을 중요하게 볼 것인가. ④ 예측 각 선택 뒤에 어떤 미래가 펼쳐질 것인가. ⑤ 가치평가 그 미래들 가운데 무엇을 더 바람직하다고 볼 것인가. ⑥ 선택 하나를 실행한다. ⑦ 세계변화 선택 때문에 상태가 변한다. ⑧ 학습 예측과 실제의 차이를 이용해 모델을 고친다. 그리고 다시 ①로 돌아갑니다. 즉: **세계 → 모델 → 선택 → 세계 → 수정된 모델** 입니다. 이 순환이:

- 바둑기사
- AI
- 동물
- 기업
- 국가
- 인간의 인생

에 모두 나타납니다. **하지만 인간에게는 AI에 없는 한 층이 더 있다** 적어도 현재의 게임 AI와 비교할 때 인간은 이상한 질문을 합니다. 나는 왜 이 게임을 하고 있지? 바둑 AI에게 승리목표는 외부에서 주어집니다. AI는: 바둑을 두는 것이 좋은가? 를 묻지 않습니다. 그러나 인간은 인생에서: 더 이기는 것이 정말 좋은가? 를 묻습니다. 돈을 많이 버는 정책이 있습니다. 하지만: 이만큼 일하면서까지? 라는 질문이 생깁니다. 명예를 크게 얻는 선택이 있습니다. 그러나: 내가 정말 원하는 삶인가? 를 묻습니다. 그러므로 인간의 자유는 단순히: **최적의 수를 찾는 능력** 만이 아닙니다. 더 깊게는: **무엇을 최적화할 것인지 다시 선택할 수 있는 능력** 에 있습니다. **아마 인생의 최고수는 '최선의 수를 많이 두는 사람'이 아닐지도 모릅니다** 한 수 한 수를 완벽하게 두려 하면 인생은 굉장히 피곤해집니다. 오히려 훌륭한 전략은:

- 치명적인 실수를 피하고
- 선택권을 오래 보존하고
- 정보가 더 생길 때까지 기다릴 줄 알고
- 작은 손실을 받아들이고
- 잘못된 길에서 빠르게 돌아오며
- 중요한 국면에 자원을 집중하는 것

일 수 있습니다. 즉 인생에서는: **최적화보다 후회 최소화** 가 더 현실적인 경우가 많습니다. 수학적으로도 이것은 별개의 문제입니다. max expected reward 와 min worst regret 은 같지 않습니다. **그래서 바둑이 우리에게 주는 가장 깊은 교훈 하나** 보통 우리는 이렇게 생각합니다. 좋은 선택을 계속하면 좋은 삶이 된다. 하지만 바둑이 보여주는 것은 조금 다릅니다. **좋은 선택은 독립적으로 존재하지 않는다.** 좋은 수는:

- 현재 판
- 상대

- 목적
- 남은 시간
- 이후의 선택
- 위험 허용도

에 따라 달라집니다. 그러므로 좋은 삶도 아마: 정답을 계속 맞히는 삶 이 아니라, **변화하는 판을 계속 제대로 읽는 삶** 에 가깝습니다. 그리고 여기서 저는 이 대화의 가장 큰 질문을 이렇게 잡고 싶습니다. **우주는 가능한 상태들을 펼친다. 생명은 그중 일부를 선택한다. 지능은 선택의 결과를 미리 상상한다. AI는 그 예측과 선택을 계산으로 증폭한다. 인간은 거기서 한 걸음 더 나아가 무엇을 선택할 가치가 있는지까지 묻는다.** 그래서 **바둑 → AI → 인생 → 우주**를 잇는 진짜 중심어는 ‘최적화’ 하나가 아니라, **가능성 · 제약 · 예측 · 선택 · 가치 · 되먹임** 이 여섯 개일 가능성이 큼니다. 그리고 이 여섯 개를 제대로 밀어붙이면 꽤 큰 독립 프로젝트가 됩니다. 다음에는 여기서 바로 “**경우의 수는 왜 지능을 낳는가 — 조합폭발에서 직관·몬테카를로·정책망·가치망·인생의 선택까지**”를 수학적으로 파고들어도 좋겠습니다.

제1부

계산할 수 있는 세계

달한 규칙 안에서 최선수를 찾는 능력과, 그 능력이 성립하기 위해 이미
주어져 있어야 하는 조건을 살핀다.

1장. 경우의 수와 지능

경우의 수가 많아지면 왜 계산량보다 선택 구조가 중요해지는가?

지능은 가능한 모든 수를 열거하는 능력이 아니라 무시할 것과 더 볼 것을 가르고, 제한된 계산을 중요한 분기에 배치하며, 결과에 따라 그 압축 방식을 수정하는 능력이다. **경우의 수는 왜 지능을 낳는가** 조합폭발 · 탐색 · 가지치기 · 정책 · 가치 · 직관 · 선택의 수학 **TARGET CLAIM:** 문제가 복잡할수록 더 많이 계산하는 존재가 더 지능적이다 **바둑판에는 361개의 점이 있다** 첫 수를 둘 때만 보면 별로 무섭지 않습니다. 361개 중 하나를 고르면 됩니다. 그러나 한 수가 놓이고 나면 다시 수백 개의 선택지가 생깁니다. 그리고 또다시 수백 개가 생깁니다. 아주 단순화하여 매 순간 평균적으로 b 개의 선택지가 있고 앞으로 d 수까지 생각한다고 해보겠습니다. 검토해야 하는 경로는 선택지 수와 탐색 깊이가 늘어날수록 폭발적으로 커집니다. 이것이 **조합폭발**입니다. 예를 들어 후보가 겨우 100개이고 10수만 앞을 보더라도 전부 검토할 수 없는 규모가 됩니다. 그러므로 중요한 사실이 하나 드러납니다. **충분히 복잡한 세계에서는 모든 미래를 계산하는 것이 원리적으로 좋은 전략이 아니다.** 시간이 먼저 끝납니다. **그렇다면 완벽한 계산기가 최강의 지능이 될 수 없다** 이것은 아주 이상한 결론입니다. 우리는 지능을 흔히: 더 많이 생각할 수 있는 능력 이라고 상상합니다. 하지만 경우의 수가 폭발하면: 계산량 $\uparrow \rightarrow$ 계산해야 할 가능성은 훨씬 더 빠르게 $\uparrow$ 합니다. 따라서 계산능력을 1,000배 늘려도 문제가 해결되지 않을 수 있습니다. 이 문제는 바둑 이전부터 체스에서도 분명했습니다. 클로드 새넨은 컴퓨터 체스를 논하면서 모든 가능한 게임 경로를 전부 조사하는 방식이 현실적으로 불가능하다는 문제를 일찍이 지적했고, 이후 게임 AI의 역사는 사실상 **전체탐색을 어떻게 포기할 것인가**의 역사였습니다. 그래서 첫 번째 역설입니다. **지능은 더 많이 보는 능력인 동시에, 무엇을 보지 않을지 결정하는 능력이다.** 바둑 **초보와 고수의 가장 큰 차이는 계산속도가 아닐 수 있다** 초보가 판을 봅니다.

가능한 수가 너무 많습니다. 여기에도 둘 수 있고, 저기에도 둘 수 있고, 잡을 수도 있고, 도망갈 수도 있습니다. 고수도 똑같은 판을 봅니다. 그런데 고수에게는 처음부터: 이 수들은 거의 아니다. 라는 것이 보입니다. 즉 초보의 문제는: **계산해야 할 후보가 너무 많다**는 것 이고, 고수의 강점은: **계산하기 전에 후보군을 줄인다**는 것 입니다. 이를 아주 추상적으로 쓰면: 361개의 가능성 $\downarrow$ 20개의 그럴듯한 수 $\downarrow$ 5개의 주요 후보 $\downarrow$ 2개의 집중 검토 $\downarrow$ 1개의 선택 입니다. 강한 지능은 계산을 시작하기 전에 이미 거대한 일을 해놓습니다. **탐색공간을 압축합니다.** 이것이 AI에서 '정책'이 하는 일이다 AlphaGo가 중요한 이유도 여기 있습니다. AlphaGo는 단순히 엄청난 컴퓨터로 바둑의 모든 경우를 계산한 것이 아닙니다. 신경망을 이용해: 이 판에서는 어떤 수들이 유망한가? 를 예측했습니다. 즉 **정책망(policy network)**이 후보수를 좁혔습니다. 그리고: 현재 이 판은 어느 쪽이 얼마나 유리한가? 를 추정하는 **가치망(value network)**을 사용했습니다. 이 둘을 몬테카를로 트리 탐색과 결합하면서 모든 경로를 동일하게 뒤지는 대신 **유망한 경로에 계산을 집중**할 수 있었습니다. AlphaGo 이후 AlphaGo Zero는 자가대국을 통해 정책과 평가를 개선하는 구조를 발전시켰습니다. KataGo는 AlphaZero형 자가대국과 신경망 유도 몬테카를로 트리 탐색을 여러 기법으로 효율화했으며, 원 논문은 비교 가능한 방식보다 계산량을 약 50배 줄인 학습 효율 개선을 보고했습니다. 따라서: **AI가 강해졌다는 것은 모든 경우의 수를 계산할 수 있게 되었다는 뜻이 아니다.** 오히려: **계산하지 않아도 될 경우의 수를 더 잘**

알아내게 되었다 에 가깝습니다. 그러면 인간의 '감'도 다시 보인다 프로기사가 판을 보자마자 말합니다. 여기 같아요. 아직 깊게 읽지 않았습니니다. 왜 그곳이 보일까요? 신비한 직감일까요? 일부는 그럴 수 있지만 훨씬 생산적인 설명은: 압축된 경험

입니다. 기사의 뇌에는 수십 년 동안:

- 모양
- 사활
- 정석
- 전투
- 집
- 두터움
- 약점
- 선후수

의 관계가 축적되어 있습니다. 그래서 처음부터 모든 점을 동등하게 보지 않습니다. 무수한 경험이: 수백 개 후보를 몇 개 후보로 압축합니다. 이때 우리는 그것을: 직관 이라고 부릅니다. 직관은 계산의 반대가 아니다 여기서 상당히 중요한 오해가 깨집니다. 우리는 흔히: 논리적 계산 ↔ 직관 을 반대로 놓습니다. 하지만 숙련자의 직관은 종종: 과거에 수행한 엄청난 학습의 압축 결과 일 수 있습니다. 즉: 반복 경험 → 패턴 발견 → 특징 압축 → 즉각적 후보 생성 입니다. 그래서 초보에게는 설명하기 어려운: 그냥 여기 느낌인데요. 가 생깁니다. '그냥'이 아닐 수 있습니다. 말로 펼치면 너무 긴 계산이 하나의 느낌으로 압축돼 있을 수 있습니다. 하지만 직관에는 치명적인 위험도 있다 압축은 정보를 버리는 행위입니다. 즉 모든 직관에는: 손실 이 있습니다. 평소에는: 이런 모양은 공격해야 한다. 가 잘 맞을 수 있습니다. 하지만 예외적인 상황에서는 틀립니다. 신진서가 KataGo와의 제2국에서 둔 흑180 빈삼각이 흥미로운 이유가 여기에 있습니다. 공식 승착은 흑160 절단이었지만, 흑180은 그가 승리를 감지한 수였습니다. 빈삼각은 일반적으로 효율이 낮은 모양이라는 지식이 있습니다. 그것은 매우 유용한 압축 규칙입니다. 하지만 실제 특정 판에서는: 일반적 모양의 아름다움보다 현재 형세의 생존이 중요할 수 있다. 즉: 휴리스틱 ≠ 진리 입니다. 그러므로 지능에는 두 층이 필요하다 첫 번째는: 빠른 압축 입니다. 모든 가능성을 검토하지 않고 좋은 후보를 만든다. 두 번째는: 압축을 의심하는 능력 입니다. 지금 이 상황은 평소 규칙의 예외인가? 를 다시 봅니다. 이를 구조화하면: 직관 → 후보 그리고 필요할 때: 의심 → 깊은 탐색 → 직관 수정 입니다. 최고수는 직관이 없는 사람이 아니라: 자신의 직관을 언제 의심해야 하는지 아는 사람 일 수 있습니다. 여기서 몬테카를로 탐색의 철학이 나온다 미래를 전부 계산할 수 없다면 어떻게 해야 할까요? 한 가지 답은:

표본을 본다 입니다. 모든 미래를 보는 대신 일부 미래를 시뮬레이션합니다. 그러면서: 이 방향으로 갔을 때 좋은 결과가 얼마나 자주 나오는가? 를 추정합니다. 바로 이 계열의 생각이 몬테카를로 방식입니다. 완전한 미래를 모릅니다. 하지만: 충분히 의미 있는 표본으로 미래의 분포를 추정한다. 이것은 굉장히 중요한 전환입니다. 완전한 앞 에서 확률적 추정 으로 이동합니다. 그래서 지능은 '미래를 아는 것'이 아니다 지능적인 존재도 미래를 모릅니다. 다만: 어떤 미래가 얼마나 그럴듯한지 를 더 잘 추정합니다. 즉: FUTURE ≠ ONE PATH 미래는: F1 F2 F3 ... 라는 가능성들의 분포입니다. 따라서 선택은: 어떤 미래가 올 것인가? 를 맞히는 문제라기보다, 가능한 미래들의 분포를 고려했을 때 어느 행동이 더 나은가? 의 문제입니다. 여기서 승률이 등장한다 어떤 수 A 이후:

- 60% 승리
- 40% 패배

라고 합시다. B 이후:

- 55% 승리
- 45% 패배

라면 단순 승패 목적에서는 A가 우월합니다. 즉: 선택 A > 선택 B 입니다. 하지만 아주 중요한 문제가 남습니다. **60%라는 숫자는 어디에서 왔는가?** 그것은 미래 전체를 직접 본 결과가 아닙니다. 모델과 탐색에서 나온 **추정치**입니다. 따라서: **평가값 $\neq$ 현실** 입니다. 평가값은 현실을 압축한 모델의 출력입니다. 이 구분이 사라지면 인간은 AI의 숫자를 신탁처럼 받아들이기 시작합니다. AlphaGo 등장 당시에도 신경망 기반 시스템이 선택한 수의 근거를 인간이 전통적 방식으로 완전히 검증하기 어렵다는 점이 중요한 문제로 제기됐습니다. **이제 첫 번째 아주 깊은 질문 최고의 선택은 가장 승률이 높은 선택인가?** 바둑에서는 상당 부분 YES입니다. 승부의 목적이 명확하기 때문입니다. 하지만 상황을 조금 바꿔봅시다. A:

- 성공하면 +10
- 실패하면 -10
- 성공확률 80%

B:

- 성공하면 +100
- 실패하면 -1
- 성공확률 30%

승률만 보면 A입니다. 기대값을 보면: B입니다. 즉: **승리확률 최대화 $\neq$ 기대가치 최대화** 입니다. 벌써 '최선'이라는 말이 흔들립니다. **더 흔들어보자** 어떤 사람이 전 재산을 걸고 있습니다. 선택 A:

- 99% 확률로 재산 10% 증가
- 1% 확률로 파산

선택 B:

- 100% 확률로 현상 유지

기대값만 따지면 A가 더 좋게 계산될 수도 있습니다. 하지만 한 번의 파산이 인생 전체를 끝낼 정도라면? **파산확률 자체가 목적함수에 들어옵니다.** 따라서 합리적 선택은: 기대수익 뿐 아니라:

- 파멸 위험
- 변동성
- 회복 가능성
- 반복 가능성
- 시간

을 포함해야 합니다. 이것이 바둑에서 '승부수'가 생기는 이유이기도 하다 현재 형세가 매우 불리합니다. 정상적인 수를 계속 두면: 95% 확률로 조금씩 저간다. 그런데 복잡한 전투를 만들면:

- 대패할 가능성 ↑
- 역전할 가능성 ↑

합니다. 승부에서는: 1집 패배와 100집 패배가 모두 패배 입니다. 따라서 뒤지고 있을 때는 분산을 키우는 전략이 합리적일 수 있습니다. 반대로 앞서고 있을 때는: 변수를 줄인다. 신진서가 KataGo와의 제3국에서 큰 전투를 피하고 안정적 흐름을 만든 사례와도 연결됩니다. 전략적으로 말하면, 복잡한 전투 분기와 인간의 실수 가능성을 줄이는 방식으로 판의 구조를 설계한 것입니다. 즉 같은 기사라도: 형세에 따라 위험 선호를 바꿉니다. 이것은 아주 중요한 일반원리다 좋은 선택은 행동 자체의 속성이 아니다. 상태와 목적의 함수다. 같은 행동이: 앞설 때는 나쁘고, 뒤질 때는 좋을 수 있습니다. 따라서: 정도로 생각해야 합니다. 여기서 인생과 연결되는 문이 열립니다. 인생에서도 위험은 위치에 따라 다르다 10억 원을 가진 사람이: 1,000만 원을 위험에 거는 것과 1,000만 원 전 재산을 가진 사람이: 1,000만 원을 거는 것은 금액은 같습니다. 그러나 전략적으로 전혀 다른 선택입니다. 즉: Risk ≠ Amount 입니다. 더 정확히는: 위험: 손실이 이후 선택공간을 얼마나 파괴하는가 라고 보는 것이 강력합니다. 여기서 새로운 개념을 도출할 수 있다 OPTION SPACE 돈 그 자체보다 중요한 것이: 이 선택 이후에도 내가 몇 개의 다음 선택을 할 수 있는가? 일 수 있습니다. A 선택은 조금 높은 수익을 주지만 실패하면 게임에서 탈락합니다. B 선택은 수익은 작지만:

- 다시 시도할 수 있고
- 정보를 얻을 수 있고
- 방향을 바꿀 수 있습니다.

그렇다면 불확실성이 매우 큰 세계에서는 B가 더 강할 수 있습니다. 이를: 선택권 보존 이라고 부를 수 있습니다. 이제 경우의 수가 많다는 것이 오히려 자산이 된다 처음에는: 선택지가 많아서 문제다. 였습니다. 그런데 이상합니다. 선택지가 하나밖에 없다면 판단은 쉽습니다. 하지만 자유도 없습니다. 선택지가 많으면 계산은 어려워집니다. 하지만: 대응 가능성도 커집니다. 따라서 경우의 수에는 두 얼굴이 있습니다. 경우의 수 ↑ → 탐색비용 ↑ 동시에 → 적응가능성 ↑ 입니다. 이것은 매우 깊은 구조입니다. 그래서 자유와 복잡성은 연결될 수 있다 아무것도 선택할 수 없는 시스템은 계산이 쉽습니다. 선택지가 무수한 시스템은 계산하기 어렵습니다. 즉 자유의 증가는: 계산 부담 을 동반할 수 있습니다. 인간에게 선택지가 많아질수록:

- 비교
- 후회
- 예측
- 정보수집
- 책임

이 증가합니다. 그러므로: 자유는 가능성의 증가이면서 동시에 최적화 문제의 증가다. AI는 이 문제를 어떻게 푸는가 AI의 핵심 전략을 아주 단순하게 압축하면: ① 후보를 만든다 무엇을 할 수 있는가. ② 유망성을 추정한다 무엇을 먼저 볼 가치가 있는가. ③ 일부 미래를 탐색한다 각 선택 뒤에 무엇이

일어날 것인가. ④ 가치를 계산한다 어느 상태가 더 좋은가. ⑤ 행동한다 ⑥ 결과를 학습한다 ⑦ 다음 판단을 개선한다 KataGo는 AlphaZero형 자기대국과 신경망 유도 몬테카를로 트리 탐색을 여러 개선으로 효율화했습니다. 원 논문은 비교 가능한 방식보다 계산량을 약 50배 줄였다고 보고합니다. 중요한 것은 단순한 연산량 확대가 아니라 **같은 거대한 상태공간에서 더 효율적으로 학습하고 탐색하는 방법**이었다는 점입니다. **인간도 거의 같은 문제를 푼다** 아침에 병원에 갑니다. 한 환자를 봅니다. 가능한 진단이 수백 개라고 해서: 의학책의 모든 질병을 하나씩 검토 하지 않습니다. 증상과 상황을 보고: 수백 개 질환 → 십여 개 → 3~5개 → 가장 가능성 높은 몇 개로 압축합니다. 그리고 검사합니다. 검사 결과가 나오면: **확률을 갱신합니다**. 즉 인간의 합리적 진단도: 후보 생성 → 확률 → 정보 획득 → 갱신 → 행동입니다. **과학도 마찬가지로** 현상을 하나 발견했습니다. 가능한 설명은 수없이 많습니다. 과학자는 모든 설명을 동시에 실험하지 않습니다. 우선:

- 기존 이론
- 관찰
- 단순성
- 메커니즘
- 선행연구

로 후보를 압축합니다. 그리고: **서로 다른 설명을 가장 잘 구별해주는 실험**을 합니다. 즉 좋은 실험은 단순히 정보를 많이 얻는 실험이 아닙니다. **가능성 공간을 가장 많이 줄이는 실험**입니다. 여기서 정보이론과 연결됩니다. **정보란 무엇인가** 놀랍게도: 모든 결과가 똑같이 예상된다. 면 정보가 적습니다. 반대로: 어떤 결과가 나오느냐에 따라 후보들이 크게 갈린다. 면 정보가 큼니다. 즉 정보는 단순한 데이터 양이 아니라: **불확실성의 감소**와 연결됩니다. 그래서 지능은: 많이 아는 것 만이 아니라 **다음에 무엇을 알아야 하는지 아는 것**입니다. 여기서 **탐색과 학습이 하나로 합쳐진다** 좋은 선택은 당장 보상을 얻는 행동일 수도 있지만, 미래 판단을 개선하기 위한 행동일 수도 있습니다. 예를 들어: A 식당은 이미 맛있는 것을 압니다. B 식당은 모릅니다. 오늘 A에 가면 만족할 확률이 높습니다. 하지만 B에 가면: 새로운 정보를 얻습니다. 이것이: **탐색 ↔ 활용** 문제입니다. 활용 = 현재 최선이라고 생각하는 것을 선택. 탐색 = 아직 잘 모르는 대안을 시험. **인생에서도 이것이 엄청나게 중요하다** 20살에는 탐색의 가치가 큼니다. 직업도, 사람도, 취향도, 자신의 능력도 잘 모르기 때문입니다. 그래서 약간의 실패를 감수하고 여러 경험을 해보는 것이: **정보 획득**이 됩니다. 반대로 어떤 영역에서는 충분히 학습한 뒤 계속 탐색하면 오히려 비용만 커질 수 있습니다. 따라서: **좋은 삶은 언제 탐색하고 언제 활용할지를 결정하는 문제이기도 하다**. 이제 **'후회'가 등장한다** 결과가 나온 뒤 사람은 생각합니다. 다른 것을 골랐으면 더 좋았을 텐데. 수학과 알고리즘에서도 **regret**, 즉 후회라는 개념을 씁니다. 대략: 최선의 선택을 계속했을 때 얻었을 보상 - 실제로 얻은 보상입니다. 중요한 전환입니다. 현실에서는 처음부터 완벽한 답을 알 수 없으므로: **완벽한 선택**보다 **시간이 지날수록 후회를 얼마나 작게 만드는가**가 더 현실적인 목표가 됩니다.

이것은 인생철학을 바꿀 수 있다 우리는 흔히: 그때 최고의 선택을 했어야 했다. 고 생각합니다. 그러나 당시에는 미래를 몰랐습니다. 따라서 더 공정한 질문은: 당시 이용할 수 있었던 정보에서 얼마나 합리적인 선택이었는가? 입니다. 결과가 나쁘다고: 선택이 나쁘다 고 바로 결론내릴 수 없습니다. 반대로 결과가 좋았다고: 판단이 훌륭했다 고 할 수도 없습니다. **HARD LOCK 01 GOOD OUTCOME ≠ GOOD DECISION** 그리고: **BAD OUTCOME ≠ BAD DECISION**입니다. 확률적

세계에서는 반드시 그렇습니다. 이제 운이라는 것이 수학적으로 들어온다 70% 성공하는 선택을 했습니다. 실패했습니다. 그렇다고 선택이 틀렸다는 뜻은 아닙니다. 30%의 세계가 실현된 것입니다. 반대로 10% 성공 전략을 택했는데 성공했습니다. 그 결과만 보고: 천재적인 전략이었다. 고 판단하면 안 됩니다. 즉 세상에는: **결정의 질** 과 **실현된 결과** 사이에 간극이 있습니다. 그 간극을 우리는 흔히: **운** 이라고 부릅니다. 그런데 반복하면 운의 힘이 달라진다 한 번의 승부에서는 약한 사람이 이길 수 있습니다. 하지만 1,000번 반복하면: **작은 확률적 우위가 누적** 됩니다. 그래서 카지노는 한 판의 승패에 크게 신경 쓰지 않습니다. 기대값이 자기에게 유리한 게임을 반복시키면 됩니다. 인생에서도: 한 번의 대박 보다 **반복 가능한 작은 우위** 가 엄청난 차이를 만들 수 있습니다. **바둑 최강자는 모든 수를 천재적으로 두는 사람이 아닐 수 있다** 매번: +100짜리 기적의 수 를 발견하는 사람보다, 매 순간: +0.2 +0.3+0.1+0.4 의 작은 우위를 거의 잃지 않는 사람이 더 강할 수 있습니다. 이 점에서 이세돌이 신진서의 특징으로 설명한: 흔들림 없이 최선에 가까운 수를 계속 둔다. 는 평가가 상당히 흥미롭습니다. 압도성은 반드시: **극적인 천재성의 빈도** 에서 나오지 않습니다. 때로는: **작은 의사결정 오류의 지속적 감소** 에서 나옵니다.

이것이 AI가 인간에게 준 가장 이상한 충격일지도 모른다 인간은 천재를:

- 영감
- 묘수
- 창조성
- 결정적 한 방

으로 기억하는 경향이 있습니다. AI는 다른 종류의 강함을 보여주었습니다. **거의 매 순간 크게 틀리지 않는다.** 작은 우위를 계속 유지합니다. 그러면 수십 수 뒤에는: 처음에는 보이지 않던 격차가 엄청나게 벌어집니다. **인생에서도 이 구조가 강할 가능성이 있다** 인생을 바꾸는 거대한 선택도 분명 있습니다. 하지만 대부분의 날은:

- 먹고
- 자고
- 운동하고
- 배우고
- 일하고
- 돈을 쓰고
- 사람에게 반응하고

작은 결정을 반복합니다. 따라서 장기적으로는: **최고의 한 수** 보다 **치명적으로 나쁜 수를 적게 두는 능력** 이 중요할 수 있습니다. 그런데 여기서 **마지막 반전이 온다** 그렇다면 AI처럼 살면 가장 좋은 인생일까요? 아닙니다. 왜냐하면 다시 목적함수 문제가 등장하기 때문입니다. 바둑 AI는: **승리하라** 라는 목적을 의심하지 않습니다. 하지만 인간은: 내가 이기려는 게임이 정말 내가 원하는 게임인가? 를 물을 수 있습니다. 승진경쟁에서 최적화할 수 있습니다. 그러나 어느 날: 나는 왜 이 조직에서 승진하려고 하지? 라고 물을 수 있습니다. 돈을 최대화할 수 있습니다. 그러나: 돈을 위해 희생하는 시간이 더 가치 있었던 것은 아닌가? 라고 물을 수 있습니다. 그래서 **인간에게는 메타 선택이 있다** 1차 선택: 무엇을 할 것인가? 2차 선택: 어떤 전략으로 할 것인가? 3차 선택: 무엇을 좋은 결과라고 정의할 것인가? 그리고 더 깊게:

4차 선택: 나는 왜 그것을 좋은 것이라고 생각하는가? 까지 올라갑니다. 여기서 바둑과 인생의 결정적 차이가 확정된다 바둑: 규칙 고정 + 목표 고정 + 판 유한 인생: 환경 변화 + 목표 변화 + 자기 자신 변화 + 타인도 전략 변화 + 종료시점 불확실 입니다. 따라서 인생은: 단순 최적화 문제가 아니다. 더 정확하게는: 최적화하면서 목적함수까지 계속 수정해야 하는 문제 입니다. 이것을 우주까지 확장하면 무엇이 남는가 바둑: 가능한 수 → 선택 → 새로운 판 AI: 상태 → 정책 → 행동 → 새로운 상태 → 학습 생명: 변이 → 환경적 차등 → 생존·번식 → 다음 세대 인간: 가능성 → 예측 → 가치판단 → 선택 → 경험 → 자기수정 이 네 세계에는 비슷한 반복구조가 있습니다. POSSIBILITY → CONSTRAINT → SELECTION → CONSEQUENCE → FEEDBACK 다만 중요한 차이가 있습니다. 바둑에서는 누군가가 목표를 부여합니다. AI에서도 인간이 학습목표와 환경을 설계합니다.

진화에는 의식적 목적을 가정할 필요가 없습니다. 인간은 스스로 목적을 반성할 수 있습니다. 그래서 현재 가장 깊게 살아남는 문장 처음에는 이렇게 생각했습니다. 세상은 거대한 최적화 문제다. 지금은 이보다 조심스럽게 수정해야 합니다. 세상에는 수많은 가능성이 있고, 제약 때문에 그중 일부만 실현되며, 어떤 시스템은 그 결과를 이용해 다음 선택을 수정한다. 그리고: 그 되먹임이 충분히 깊어질 때 우리가 '학습'과 '지능'을 보기 시작한다. 아직 이것을 우주의 보편법칙이라고 부를 단계는 아닙니다. 그러나 바둑·AI·생명·인간을 연결하는 매우 강한 후보 구조입니다. MASTER MODEL v0.2 THE FINITE MIND IN AN EXPLODING WORLD 세계가 제공하는 가능성은 ↓ 계산능력보다 빠르게 증가한다. 따라서 유한한 지능은 ↓ 모든 가능성을 볼 수 없다. 그러므로 ↓ 중요한 가능성을 선별한다. 선별하려면 ↓ 세계를 압축한 모델이 필요하다. 모델을 이용해 ↓ 미래를 예측한다. 예측된 미래를 ↓ 가치에 따라 평가한다. 하나를 선택하면 ↓ 세계가 변한다. 실제 결과와 예측의 차이가 ↓ 모델을 수정한다. 그리고 다시 시작한다. 세계 → 압축 → 예측 → 선택 → 결과 → 학습 → 새로운 세계 이 노드가 열어놓은 다음 질문 그러나 아직 가장 어려운 문제가 남아 있습니다. 우리는 계속: 더 좋은 미래 를 말했습니다. 그런데 도대체 '좋다'는 값을 누가 정합니까? 바둑에서는 승리입니다. AI에서는 보상함수입니다. 진화에서는 생존·번식의 차이가 결과를 걸러냅니다. 하지만 인간에게는:

- 행복
- 돈
- 자유
- 사랑
- 권력
- 지식
- 명예
- 의미
- 도덕

이 충돌합니다. 따라서 다음 노드에서 들어가야 할 곳은 경우의 수가 아닙니다. 가치함수는 어디에서 오는가 입니다. 그리고 이것은 바둑에서 인생으로 넘어가는 진짜 경계입니다. 계산은 무엇이 가능한지를 말해줄 수 있다. 확률은 무엇이 일어날 법한지를 말해줄 수 있다. 그러나 그중 무엇이 일어나야 하는지는 계산만으로 나오지 않는다. 다음은 여기서 “AI는 승률을 계산할 수 있는데 왜 삶의 목적을 계산할 수 없는가 — 목적함수·보상·효용·다목적 최적화·도덕·의미”로 넘어가는 것이 가장 자연스럽습니다.

2장. 목표와 불확실성

바둑에는 승률이 있는데 왜 인생에는 하나의 정답 승률이 없는가?

승률과 최적성은 규칙·목표·상태표현이 이미 정해진 문제 안에서만 성립한다. 인생에서는 여러 가치가 충돌하고 목표가 변하며 결과의 의미도 행위 뒤에 다시 해석된다. AI는 승률을 계산할 수 있는데 왜 삶의 목적은 계산할 수 없는가 목적함수 · 보상 · 효용 · 다목적 최적화 · 가치충돌 · 의미 TARGET CLAIM: 충분히 좋은 예측모델이 있으면 인간의 최선의 삶도 계산할 수 있다 바둑 AI는 왜 그렇게 강한가 이유는 의외로 단순합니다. 바둑의 최종 목표가 비교적 명확하기 때문입니다. 이기면 된다. 판이 끝났을 때:

- 승
- 패

를 판정할 수 있습니다. 그러므로 AI는 엄청나게 복잡한 중간 상황도 결국 하나의 축으로 환원할 수 있습니다. 이 판은 승리 가능성을 얼마나 높이는가? 즉 수많은 수를 하나의 공통 척도로 비교할 수 있습니다. 이것이 최적화의 출발점입니다. 최적화에는 반드시 '좋다'의 정의가 필요하다 어떤 선택지들이 있다고 합시다. A, B, C. AI에게: 가장 좋은 것을 골라. 라고 하면 아무 의미가 없습니다. 무엇이 좋은지 정의하지 않았기 때문입니다. 예를 들어:

- 가장 빠른 것
- 가장 싼 것
- 가장 안전한 것
- 가장 아름다운 것

가운데 무엇을 최대화할지 정해야 합니다. 즉 모든 최적화 뒤에는: 목적함수 가 숨어 있습니다. 수학적으로는 매우 단순하게: maximize $U(x)$ 라고 쓸 수 있습니다. 여기서 U는: 얼마나 좋은지를 숫자로 평가하는 함수 입니다. 그런데 인간의 삶에는 '승리'가 없다 이제 문제가 생깁니다. 사람 A가 있습니다. 선택 1:

- 연봉 5억
- 주 80시간 근무
- 높은 명예
- 가족과 거의 못 만남

선택 2:

- 연봉 1억
- 주 35시간 근무
- 가족과 많은 시간

- 사회적 지위는 낮음

어느 쪽이 더 좋은 삶입니까? 계산하려면 먼저 물어야 합니다. 돈과 시간은 얼마나 교환 가능한가? 가족관계와 명예는? 자유와 안정은? 이미 문제는 숫자가 아닙니다. 서로 다른 가치의 비교입니다. 삶은 다목적 최적화 문제다 인간은 동시에 여러 것을 원합니다. 예를 들어:

- 건강
- 돈
- 자유
- 관계
- 성취
- 즐거움
- 지식
- 안전
- 사회적 인정
- 의미

따라서 삶의 여러 가치를 억지로 하나의 점수로 표현하더라도 문제가 남습니다. 문제는: 각 항목의 가중치를 모른다. 더 큰 문제는: 그 가중치가 사람마다 다르다. 그리고 더 큰 문제는: 같은 사람 안에서도 계속 바뀐다. 20살의 목적함수와 70살의 목적함수는 같지 않을 수 있다 20살에는:

- 도전
- 성장
- 새로움
- 가능성

이 중요할 수 있습니다. 50살에는:

- 안정
- 가족
- 영향력
- 건강

이 커질 수 있습니다. 80살에는:

- 고통 감소
- 존엄
- 사랑하는 사람
- 기억
- 삶의 정리

가 중요해질 수 있습니다. 즉 인간은: 고정된 효용함수를 가지고 태어나 평생 최적화하는 기계가 아닙니다. 삶을 살아가면서: 효용함수 자체가 변합니다. 이것이 아주 중요한 되먹임이다 앞 장에서는: 선택 → 세계 변화 → 학습 을 봤습니다. 여기에 한 층을 추가해야 합니다. 선택 → 경험 → 자기 변화 →

가치 변화입니다. 따라서 인간은 단순히 세계를 학습하는 것이 아닙니다. 세계를 살아가면서 무엇을 중요하게 여기는 존재인지도 바뀐다. 그러므로 인간의 선택은 자기 자신을 선택하는 행위이기도 하다. 직업을 선택합니다. 표면적으로는: 어떤 일을 할 것인가? 입니다. 그러나 20년 반복하면:

- 사용하는 언어
- 만나는 사람
- 사고방식
- 시간감각
- 욕망
- 능력
- 가치

가 바뀝니다. 따라서 장기적 선택은: 결과를 고르는 것 이면서 동시에 미래의 나를 만드는 것 입니다. 이것이 바둑에는 거의 없는 층입니다. 바둑들은 자신이 어떤 돌이 되고 싶은지 고민하지 않는다 이 문장은 우스워 보이지만 중요합니다. 바둑의 상태는 변합니다. 그러나: 승리라는 목적 자체 는 변하지 않습니다. 인간은 다릅니다. 우리는 어떤 목표를 추구하다가: 이 목표를 계속 추구하고 싶은 사람이 되고 있는가? 를 물을 수 있습니다. 즉: 선택자 자체가 선택 과정의 산출물이다. 여기서 파레토 최적이라는 개념이 중요해진다 다목적 최적화에서 한 해가 완벽하게 최고가 아닐 수 있습니다. 예: A:

- 돈 10
- 자유 2

B:

- 돈 7
- 자유 7

C:

- 돈 3
- 자유 10

누가 최고입니까? 가중치를 모르면 답할 수 없습니다. 이때 중요한 것은: 다른 어떤 선택도 모든 면에서 더 좋지 않은 선택들 입니다. 이를 파레토 최적이라고 부릅니다. 즉 A, B, C 모두 살아남을 수 있습니다. 이것은 삶에 상당히 잘 맞는다 우리는 흔히: 최고의 직업 최고의 도시 최고의 배우자 최고의 삶 을 찾습니다. 하지만 실제로는: 서로 다른 가치 조합의 여러 좋은 해 가 존재할 수 있습니다. 서울의 삶. 제주의 삶. 대기업의 삶. 작은 사업의 삶. 학자의 삶. 예술가의 삶. 어느 하나가 모든 사람에게 절대적으로 우월하지 않을 수 있습니다. 즉: BEST LIFE보다 PARETO FRONTIER 가 더 적절한 그림일 수 있습니다. 그런데 사람은 결국 하나를 골라야 한다 파레토 전선에 좋은 선택지가 100개 있습니다. 그래도 우리는 하나의 삶을 살아야 합니다. 결국: 가치 간 교환비율 을 정해야 합니다. 예를 들어: 돈 1억을 더 벌기 위해 자유시간을 얼마까지 줄일 것인가? 안정성을 위해 얼마만큼의 성장 가능성을 포기할 것인가? 이때 수학은 선택구조를 보여줄 수 있습니다. 하지만: 교환비율의 정당성까지 자동으로 결정하지 못합니다. AI도 여기에서 멈춘다 아무리 완벽한 예측 AI가 있다고 해봅시다. AI가

말합니다. 선택 A를 하면:

- 20년 후 자산 30억
- 이혼확률 60%
- 건강수명 75세
- 사회적 영향력 높음

선택 B를 하면:

- 자산 8억
- 이혼확률 15%
- 건강수명 82세
- 사회적 영향력 중간

완벽하게 미래를 예측했다고 합시다. 그래도 질문이 남습니다. **어느 것이 더 좋은가?** 예측은 가치판단을 제거하지 못합니다. 오히려 더 선명하게 드러냅니다. 이것이 ‘정렬’ 문제의 철학적 핵심이다 AI 분야에서 중요한 문제 가운데 하나가: **alignment** 입니다. 즉: AI가 무엇을 잘하게 만들 것인가? 입니다. AI가 목표를 매우 잘 달성하는 것은 오히려 위험할 수 있습니다. 목표가 잘못 정의됐다면: **잘못된 목적을 극도로 효율적으로 달성** 하기 때문입니다. 이것은 최적화의 가장 무서운 역설입니다. **측정 가능한 것이 목적이 되는 순간** 조직에서도 흔히 일어납니다. 학교의 목표: 잘 배우게 한다. 측정지표: 시험점수. 그러면 시스템은 점수를 최적화하기 시작합니다. 병원의 목표: 환자를 잘 치료한다. 지표: 환자 수, 매출, 평균 대기시간. 그러면 지표가 실제 목적을 잠식할 수 있습니다. 회사의 목표: 장기적 가치 창출. 지표: 분기 실적. 그러면 장기투자를 줄일 수 있습니다. 이것이 흔히 **굿하트의 법칙**과 연결되는 문제입니다. 측정값이 목표가 되면 좋은 측정값으로서의 기능을 잃기 쉽다. 이것이 인생 최적화의 가장 큰 함정 중 하나다 사람은 숫자를 좋아합니다.

- 연봉
- 자산
- 팔로워
- 조회수
- 논문 수
- 성적
- 매출

숫자는 비교가 쉽습니다. 그래서 최적화하기 쉽습니다. 반면:

- 좋은 관계
- 지적 자유
- 존엄
- 만족
- 의미

는 측정하기 어렵습니다. 그러므로 이상한 일이 벌어집니다. **중요한 것보다 측정하기 쉬운 것을 더**

열심히 최적화한다. AI 시대에는 이것이 더 강해질 수 있다 AI는 측정 가능한 목적을 최적화하는 데 매우 강합니다. 그러므로 조직이: 고객 만족도를 높여. 라고 말하면서 실제로는 클릭률을 보상하면, AI는 클릭률을 최적화합니다. 그리고 클릭을 늘리기 위해:

- 자극
- 분노
- 중독적 반복
- 극단적 콘텐츠

를 활용할 수도 있습니다. 즉: **Objective specification matters.** 목적을 어떻게 번역했는지가 결과를 지배합니다. **그러면 인간은 정말 자신의 목적을 알고 있을까** 여기서 상황은 더 복잡해집니다. AI에게 좋은 목적함수를 주면 된다고 생각할 수 있습니다. 그런데 인간 자신도: 나는 무엇을 원하는가? 를 정확히 모르는 경우가 많습니다. 우리는:

- 다른 사람의 욕망을 모방하고
- 사회적 지위를 욕망하고
- 광고에 영향받고
- 경험하기 전에는 선호를 모르고
- 얻고 나서야 원하지 않았다는 것을 깨닫습니다.

따라서: **preference $\neq$ fixed input** 입니다. 선호는 만들어지고 바뀝니다. **이때 고전적인 의사결정 이론이 흔들린다** 보통 선택은: 현재 선호 + 미래 결과 예측 으로 한다고 봅니다. 그런데 미래의 내가 다른 선호를 가진다면? 현재의 나는: 미래의 나를 대신해서 결정합니다. 그러면 질문이 생깁니다. **어느 시점의 나의 선호가 진짜 나의 선호인가?** 쉽지 않습니다. 이것은 인생을 '게임'이라고 부를 때 생기는 **한계**다 게임에서는:

- 플레이어
- 규칙
- 목표

가 구분됩니다. 인생에서는: **플레이어도 변한다.** 즉: $\text{Player}(t_0) \neq \text{Player}(t_{20})$ 일 수 있습니다. 20년 뒤 나는:

- 다른 기억
- 다른 몸
- 다른 관계
- 다른 가치

를 가질 수 있습니다. 그래서 현재의 선택은 사실: **미래의 나에게 영향을 미치는 선택** 이기도 합니다. **그러면 삶의 좋은 전략은 목적함수를 고정하는 것일까** 아마 아닐 가능성이 큼니다. 오히려: **목적함수를 수정할 수 있는 능력** 이 중요할 수 있습니다. 하지만 이것도 위험합니다. 매번 가치가 바뀌면 아무것도 지속할 수 없습니다. 따라서 두 힘이 필요합니다. 안정성 어떤 가치는 충분히 오래 유지한다. 수정가능성 새로운 경험과 증거가 쌓이면 바꾼다. 즉: **가치에도 exploration $\leftrightarrow$ exploitation** 문제가

있다. 이것을 AI 언어로 번역하면 재미있다 일반 AI: state → policy → action → reward 인간: state → self-model → values → policy → action → experience → self-change → value-update 즉 인간에서는: 보상함수가 루프 안에 들어옵니다. 이것은 구조적으로 엄청난 차이입니다. 그렇다면 의미는 무엇인가 여기서 '의미'라는 말을 함부로 쓰면 안 됩니다. 우선 최소한으로 정의해봅시다. 의미는 단순한 즐거움과 다릅니다. 어떤 행동이:

- 내가 중요하다고 여기는 것과 연결되고
- 장기적인 자기서사에 들어가며
- 나를 넘어서는 관계나 목표와 연결될 때

사람은 그것을 의미 있다고 경험할 수 있습니다. 즉 의미는 대략: **경험이 가치체계와 연결되는 방식**으로 생각할 수 있습니다. 그래서 고통과 의미가 공존할 수 있다 달리기 훈련은 힘듭니다. 육아도 힘듭니다. 연구도 힘듭니다. 환자를 돌보는 것도 힘듭니다. 하지만: 중요하다고 생각하는 목표와 연결되면 사람은 고통을 감수합니다. 즉 단순 행복 최대화 모델에서는 설명하기 어려운 선택이 나타납니다. 인간은 항상 쾌락을 최대화하지 않는다. 여기서 도덕이 등장한다 지금까지는: 나에게 좋은 것 을 이야기했습니다. 그런데 인간은: 타인에게 어떤 영향을 주는가? 도 고려합니다. 즉 목적함수에 다른 사람이 들어옵니다. 예: 내 이익 +100 타인의 피해 -1000 이라면? 단순 자기효용 최대화로는 충분하지 않습니다. 여기서 윤리학이 시작됩니다. 윤리학은 목적함수 설계 문제인가 어느 정도는 그렇다고 볼 수 있습니다. 공리주의는 대략: **전체 복지를 크게 하라** 는 형태입니다. 의무론은: 어떤 행동은 결과와 무관하게 금지되거나 요구된다. 는 제약을 둡니다. 덕윤리는: 어떤 사람이 되어야 하는가? 에 더 관심을 둡니다. 즉 세 윤리전통을 아주 거칠게 번역하면:

- 공리주의 → 목적함수
- 의무론 → 제약조건
- 덕윤리 → 정책을 생성하는 행위자의 성향

으로 볼 수 있습니다. 물론 완전히 같지는 않지만 비교틀로는 꽤 흥미롭습니다. 이것은 바둑보다 훨씬 복잡한 최적화다 바둑에서는: maximize win probability 이면 됩니다. 윤리적 삶에서는: maximize welfare 하면서 rights constraints 를 지키고, 동시에 character formation 까지 고려할 수 있습니다. 즉: 목적 + 제약 + 행위자 변화 가 한꺼번에 들어옵니다. 그렇다고 AI가 가치에 아무 도움도 못 준다는 뜻은 아니다 중요한 구분입니다. AI가: 가치를 결정할 수 없다 와 가치판단에 도움을 줄 수 없다 는 전혀 다릅니다. AI는:

- 내가 실제 행동에서 무엇을 반복해서 선택하는지 보여주고
- 서로 모순되는 가치들을 지적하고
- 장기 결과를 시뮬레이션하고
- 숨겨진 이해관계자를 찾고
- 다른 윤리적 관점에서 결과를 비교

할 수 있습니다. 즉 AI는: 가치의 판정자 가 아니라 가치의 거울과 시뮬레이터 가 될 수 있습니다. 이제 전체 프로젝트에서 중요한 분리가 하나 생긴다 예측 문제 무엇이 일어날 것인가? AI가 점점 강해질 수 있습니다. 최적화 문제 목표가 주어졌을 때 무엇을 해야 하는가? AI가 매우 강해질 수 있습니다.

가치 문제 무엇을 목표로 삼아야 하는가? 여기서는 인간·사회·철학의 문제가 남습니다. 메타가치 문제 가치체계를 언제 왜 수정해야 하는가? 이것은 더 깊습니다. 그리고 여기서 우주로 넘어갈 때 반드시 **멈춰야 한다** 우주에도:

- 안정
- 선택
- 경쟁
- 최소화
- 극값
- 되먹임

같은 구조가 보입니다. 그래서: 우주도 어떤 목적함수를 최적화하는가? 라는 유혹이 생깁니다. 하지만 아직 근거가 없습니다. 물리학에서 어떤 과정이 최소작용원리 등으로 기술된다고 해서: **우주가 '목적'을 가지고 최솟값을 선택한다**고 해석할 필요는 없습니다. 수학적 최적화 표현과 목적론은 구별해야 합니다. **HARD LOCK 02 MATHEMATICAL EXTREMUM ≠ COSMIC PURPOSE** 입니다. 이 구분을 잃으면 과학에서 형이상학으로 몰래 넘어갑니다. **현재까지의 통합모델을 갱신한다 MASTER MODEL v0.3 LEVEL 1 — POSSIBILITY** 무엇이 가능한가. **LEVEL 2 — CONSTRAINT** 무엇이 불가능한가. **LEVEL 3 — MODEL** 세계의 구조를 어떻게 압축하는가. **LEVEL 4 — PREDICTION** 각 선택 뒤에 무엇이 일어날 것인가. **LEVEL 5 — VALUE** 어떤 결과를 더 좋은 것으로 볼 것인가. **LEVEL 6 — SELECTION** 무엇을 실제로 할 것인가. **LEVEL 7 — CONSEQUENCE** 세계와 자신이 어떻게 바뀌는가. **LEVEL 8 — LEARNING** 예측과 실제의 차이로 모델을 수정한다. **LEVEL 9 — VALUE REVISION** 경험을 통해 무엇을 중요하게 여길지도 수정한다. 그리고 다시 시작합니다. **세계 → 모델 → 미래 → 가치 → 선택 → 결과 → 학습 → 자기변화 이 프로젝트의 질문도 한 단계 깊어졌다** 처음 질문: 어떻게 최선의 선택을 할 것인가? 앞 장 이후: 불확실한 세계에서 어떻게 좋은 선택을 할 것인가? 앞 장 이후에는 질문 자체가 변합니다. **무엇을 '좋다'고 정의해야 하는가?** 그리고 더 깊은 질문: **그 정의를 우리는 어디서 얻는가?**

- 진화?
- 욕망?
- 문화?
- 이성?
- 사회?
- 경험?
- 도덕?
- 자기서사?

여기서부터 바둑과 AI만으로는 답할 수 없습니다. 철학, 진화생물학, 신경과학, 경제학, 윤리학이 한꺼번에 들어와야 합니다. 바둑판에는 승률이 있는데 왜 인생에는 '정답 승률'이 없는가 **확률 · 위험 · 불확실성 · 무지 · 모델오류 · 강건성 · 블랙스완 TARGET CLAIM:** 좋은 의사결정은 각 선택지의 성공확률을 정확히 계산해서 가장 높은 것을 고르는 것이다 **우리는 지금까지 너무 쉽게 '60%'라고 말했다** 바둑 중계를 보면 이런 숫자가 뜹니다. 흑 승률 63.7% 그러면 마치 미래의 63.7%가 어딘가에 실제로 존재하는 것처럼 느껴집니다. 하지만 먼저 질문해야 합니다. **63.7%는 어디에 존재하는가?**

바둑판 위에 적혀 있습니까? 아닙니다. 자연법칙에 새겨져 있습니까? 아닙니다. 그 숫자는:

- 현재 판
- 사용하는 엔진과 모델
- 적용한 규칙과 덤
- 탐색량과 탐색 설정

를 넣어 만들어낸 평가값입니다. 즉: 승률은 미래 그 자체가 아니라 특정 계산 조건 아래 나온 모델의 판단이다. 그런데 사업을 시작한다고 해보자 새로운 AI 회사를 만듭니다. 질문합니다. 5년 안에 성공할 확률은? 누군가 말합니다. 23%. 여기서 바로 문제가 시작됩니다. 23%라는 숫자의 근거는 무엇입니까?

- 과거 스타트업 통계?
- 같은 산업?
- 같은 창업자?
- 같은 금리환경?
- 같은 AI 기술수준?
- 같은 경쟁자?
- 앞으로 등장할 규제?

완전히 같은 실험을 수천 번 반복할 수도 없습니다. 따라서 주사위의 1/6과 스타트업의 23%는: 같은 종류의 확률이 아닐 수 있다. 확률을 모른다는 것은 두 종류가 있다 A. 숫자를 모른다 예: 성공확률이 30%인지 40%인지 모르겠다. 그래도:

- 가능한 사건
- 인과구조
- 주요 변수

는 대략 알고 있습니다. B. 어떤 숫자를 써야 하는지조차 모른다 예: 5년 뒤 AI 산업이 어떤 구조인지 모르겠다. 그러면 성공확률 30%라는 숫자 자체가 이상해집니다. 더 깊게는: 무엇을 모르고 있는지 모를 수도 있다. 여기서 현실세계가 바둑판과 완전히 갈라집니다. 이제 아주 중요한 차이가 생긴다 계산하기 어려운 문제와 계산할 숫자 자체가 없는 문제는 다릅니다. 바둑은 주로 첫 번째입니다. 현실의 상당 부분은 두 번째까지 갑니다. 이것을 혼동하면: AI가 충분히 강해지면 모든 결정을 계산할 수 있다. 는 착각이 생깁니다. 계산능력이 무한해져도: 잘못된 모델을 넣으면 정밀하게 틀립니다. 이것을 수학적으로 아주 단순하게 보자 우리는 기대값을: $EV = \sum p_i \times u_i$ 처럼 계산합니다. 여기서: p_i = 결과 i 의 확률 u_i = 결과 i 의 가치입니다. 아주 강력한 공식입니다. 하지만 잘 보면 두 가지를 이미 요구합니다.

1. 가능한 결과를 알아야 한다.
2. 각 결과의 확률을 알아야 한다.

앞 장에서 이미:

3. 결과의 가치도 알아야 한다.

가 추가되었습니다. 즉 기대효용 계산에는 사실: 세계모델 + 확률모델 + 가치모델 세 가지가 들어갑니다.

이것이 블랙스완의 핵심 문제다 '블랙스완'을 단순히: 아주 드문 사건 이라고 이해하면 부족합니다. 더 중요한 것은: 기존 모델이 중요하게 고려하지 않았던 사건이 큰 결과를 만드는 것 입니다. 즉: 낮은 확률 보다 더 깊은 문제는: 모델 밖 입니다. 우리가 확률을 0.1%라고 계산했다면 이미 사건을 알고 있습니다. 진짜 어려운 사건은: 후보목록 자체에 없었다. 입니다. 인생은 특히 그렇다 20살의 사람이 40년 계획을 세운다고 해봅시다. 그는:

- AI
- 스마트폰
- 특정 질병
- 새로운 인간관계
- 경제위기
- 자기 가치관 변화

를 정확히 예상하지 못합니다. 그래서: 40년 뒤 가장 좋은 상태를 계산해서 오늘 최적경로를 선택한다. 는 모델에는 한계가 있습니다. 장기 예측이 어려워질수록 전략이 변해야 합니다. 예측이 약해지면 무엇이 강해져야 하는가 여기서 아주 중요한 전환이 생깁니다. OPTIMALITY → ROBUSTNESS 입니다. 최적화는 묻습니다. 가장 가능성이 높은 미래에서 무엇이 최고인가? 강건성은 묻습니다. 내가 틀렸더라도 어느 선택이 여러 미래에서 견디는가? RAND의 강건한 의사결정 접근도 단일한 '가장 그럴듯한 미래'를 정확히 예측해 그 미래에 최적화하기보다, 다양한 가능한 미래에서 받아들일 만한 성과를 내는 정책을 찾는 방향을 강조합니다. 여기서 인생의 새로운 목적함수가 나타난다 기존: 가장 좋은 미래를 만든다. 수정: 여러 미래에서 치명적으로 망하지 않는다. 이것은 소극적으로 들립니다. 그러나 불확실성이 큰 환경에서는 굉장히 강력합니다. 왜냐하면: 살아남은 사람에게만 다음 선택이 있기 때문입니다. 앞 장의 선택권 보존이 다시 등장한다 우리가 앞서 말했습니다. 좋은 선택은 미래의 선택공간을 보존할 수 있다. 이제 이유가 분명해집니다. 미래를 정확히 모르면: 미래 정보를 받은 뒤 다시 선택할 권리 가 매우 가치 있습니다. 따라서:

- 현금
- 건강
- 다양한 기술
- 좋은 인간관계
- 낮은 고정비
- 이동 가능성
- 학습능력

같은 자원은 단순히 현재의 효용만 주는 것이 아닙니다. 미래의 선택권 을 줍니다. 그래서 '최적 계획'보다 '적응 가능한 계획'이 강할 수 있다 예: A 인생설계: 25세 입사 → 30세 승진 → 35세 집 → 40세 임원 → 55세 은퇴 모든 것이 맞으면 효율적입니다. 하지만 구조가 하나만 어긋나도 전체 경로가 흔들릴 수 있습니다. 다른 설계: 방향 설정 → 작은 실험 → 정보 획득 → 수정 → 다음 선택 입니다. 이것은 처음부터 완벽한 경로를 알지 못합니다. 대신: 계속 다시 계산한다. 의학적 진단을 생각하면 아주 분명하다 환자가 왔습니다. 진단: A 50% B 30% C 20% 라고 생각합니다. 그런데 검사 하나를 하면: A와 B를 거의 구별할 수 있습니다. 그러면 바로 치료하는 것보다: 먼저 검사하는 것이 최적 일 수 있습니다. 왜냐하면 검사가 직접 건강을 좋게 하지는 않지만: 다음 결정의 질을 높이기 때문입니다.

따라서 행동의 가치에는 두 층이 있다 직접 가치 그 행동 자체가 주는 결과. 정보 가치 그 행동을 통해 미래의 판단이 좋아지는 정도. 이것을 인생에 가져오면 놀랍습니다. 첫 직업을 선택할 때: 연봉이 얼마인가? 만 볼 것이 아니라: 이 경험이 나 자신과 세상에 대해 얼마나 많은 정보를 줄 것인가? 도 볼 수 있습니다. 그래서 좋은 탐색에는 조건이 있다 아무 실패나 가치 있는 것은 아닙니다. 좋은 실험은: 실패해도 죽지 않아야 한다. 이것이 중요합니다. 작게 실패하고, 많이 배우고, 다시 시도할 수 있어야 합니다. 즉: downside 제한 + learning potential ↑ + repeatability ↑ 가 강력합니다. 선택을 되돌릴 수 있느냐가 핵심 변수가 된다 REVERSIBLE 틀리면 돌아올 수 있다. 이 경우: 빨리 시험하는 것 의 가치가 큼니다. IRREVERSIBLE 틀리면 회복이 어렵다. 이 경우: 더 많은 정보와 안전마진 이 필요합니다. 따라서 좋은 의사결정은:

모든 결정에 똑같이 오래 고민하는 것이 아니다. 결정의 되돌림 가능성에 따라 사고비용을 배분합니다. 이것은 계산자원의 배분 문제다 앞 장로 다시 돌아옵니다. 우리에게는 사고시간도 유한합니다. 따라서:

- 작은 결정
- 되돌릴 수 있는 결정

에 10시간씩 쓰면 비효율적입니다. 반대로:

- 치명적
- 비가역적
- 장기적

결정을 10초 만에 하면 위험합니다. 즉: 생각하는 시간 자체도 최적화 대상 입니다. 여기에 한 변수를 더 넣을 수 있다 INFORMATION GAIN 즉: 더 생각하거나 조사하면 실제로 판단이 얼마나 좋아지는가? 입니다. 정보를 아무리 더 모아도 결론이 거의 안 바뀐다면: 그만 생각해야 합니다. 반대로 작은 검사 하나가 판단을 완전히 바꿀 수 있다면: 먼저 정보를 얻어야 합니다. 그래서 결정에는 '멈춤 규칙'도 필요하다 끝없이 정보만 모으는 것도 실패입니다. 어느 순간에는: 충분히 알았다. 고 결정해야 합니다. 즉: 추가정보의 기대가치 < 정보 획득 비용 이 되는 순간 행동합니다. 완벽한 확신까지 기다리는 것이 아닙니다. 충분한 확신 에서 움직입니다. 인간의 불안은 종종 이 멈춤 실패와 관계있다 혹시 이것도? 이것까지 확인해야 하지 않을까? 다른 가능성은? 가능성 공간은 끝없이 펼쳐집니다. 그러므로 완전한 확실성을 요구하면: 행동할 수 없습니다. 현실의 합리성은: 불확실성이 없어질 때까지 기다리는 것 이 아니라 남아 있는 불확실성을 안고 행동하는 능력 까지 포함합니다. 베이지적 사고가 여기서 강하다 베이지적 접근의 핵심 직관: 새로운 증거가 들어오면 믿음의 정도를 갱신한다. 입니다. 처음부터 절대적으로 맞는 답을 찾기보다: 사전 믿음 + 새 증거 ↓ 수정된 믿음 으로 갑니다. 베이지 인식론과 베이지 결정이론은 확률을 합리적 믿음의 정도와 연결해 이 갱신 문제를 체계화합니다. 하지만 베이지도 마법은 아니다 베이지 공식은: 어떤 모델도 필요 없다. 는 뜻이 아닙니다. 여전히:

- 가능한 가설
- 사전확률
- 데이터 생성구조

가 필요합니다. 즉: Bayesian updating cannot update a hypothesis you never imagined. 후보 자체에 없는 사건은 갱신하기 어렵습니다. 그래서 열린 세계에서는: 새로운 가설을 만드는 능력

까지 필요합니다. **최고의 바둑수도 후보군 밖에서 나타날 수 있다** 인간이: 이 모양에서는 A 아니면 B. 라고 생각할 때 AI가: C. 를 둡니다. 처음에는 악수처럼 보입니다. 그런데 분석하면 좋은 수입니다. AI 혁명이 바둑계에 준 중요한 충격 중 하나는 바로: **인간이 생각하지 않던 후보의 발견** 이었습니다. 즉 AI는 단순히: 인간 후보 중 최선을 잘 골랐다. 가 아니라, **후보공간 자체를 확장** 하기도 했습니다. 이제 **지능의 정의가 다시 확장된다** 앞 장: **탐색공간을 잘 줄이는 능력** 앞 장: **무엇을 가치 있게 볼지 반성하는 능력** 앞 장: **불확실성을 다루면서 새로운 선택지를 만들고 수정하는 능력** 까지 옵니다. 따라서 지능은 단순한: **OPTIMIZATION** 보다 훨씬 넓습니다. **현재의 인간 의사결정을 하나의 루프로 압축하면** ① **WORLD** 무슨 상황인가. ↓ ② **MODEL** 무엇이 일어나고 있다고 생각하는가. ↓ ③ **UNCERTAINTY** 무엇을 얼마나 모르는가. ↓ ④ **OPTION GENERATION** 무엇을 할 수 있는가. ↓ ⑤ **PREDICTION** 각 선택은 어떤 미래를 만들 것인가. ↓ ⑥ **VALUE** 어떤 미래를 원하는가. ↓ ⑦ **ROBUSTNESS** 내 예측이 틀려도 견딜 수 있는가. ↓ ⑧ **ACTION** 행동한다. ↓ ⑨ **INFORMATION** 무엇을 새롭게 배웠는가. ↓ ⑩ **UPDATE** 세계모델·자기모델·가치를 수정한다. 그리고 다시 시작합니다. **이것이 바둑과 인생을 잇는 가장 아름다운 비유일지도 모른다** 바둑기사는: 좋은 판을 달라. 고 요구할 수 없습니다. 현재 판에서 뒤야 합니다. 인간도: 완벽한 조건이 오면 살겠다. 고 할 수 없습니다. 현재:

- 몸
- 돈
- 관계
- 시대
- 능력

에서 시작해야 합니다. 하지만: **지금의 상태는 다음 수를 제약할 뿐, 다음 수를 완전히 결정하지는 않는다.** 이제 **우주의 문제로 한 층 더 간다** 여기서 정말 이상한 질문이 생깁니다. 바둑에서: 가능한 수들 가운데 하나가 실현됩니다. 인생에서도: 가능한 행동들 가운데 하나가 선택됩니다. 양자역학에서는: 여러 가능한 측정결과와 확률구조가 등장합니다. 그렇다면: 자연 자체도 가능성들 가운데 하나를 '선택'하는 것인가? 이 질문은 매우 유혹적입니다. 하지만 여기서부터는 과학적 사실과 철학적 해석을 극도로 조심해서 분리해야 합니다. **우리가 지금까지 말한 '선택'에는 사실 세 종류가 있었다** 선택 1 — 인간의 선택 의식적 판단이 있을 수 있습니다. 선택 2 — 자연선택 의식적 선택자가 필요하지 않습니다. 차등 생존·번식의 결과입니다. 선택 3 — **물리적 결과의 실현** 어떤 상태에서 특정 결과가 발생합니다. 이것을 인간 선택과 같은 의미의: 자연의 선택 이라고 부르면 위험합니다. **HARD LOCK 03 OUT-COME SELECTION ≠ INTENTIONAL CHOICE** 입니다. 'selection'이라는 같은 단어가 구조적 유사성을 만들어내지만 메커니즘은 다를 수 있습니다. **지금까지의 MASTER를 한 문장으로 다시 쓴다** 앞 장에서는: **세계는 가능성을 만들고 지능은 선택한다.** 라고 생각했습니다. 지금은 더 정교하게 수정할 수 있습니다. **유한한 행위자는 완전히 알 수 없는 세계에서 제한된 모델을 만들고, 가능한 행동 일부를 상상하며, 불완전한 가치판단 아래 하나를 선택하고, 그 결과를 이용해 세계와 자기 자신에 대한 모델을 계속 수정한다.** 이것이 현재까지 가장 강한 인간 의사결정 모델입니다. 그리고 여기서 **신진서의 빈삼각이 다시 등장한다** 제2국의 공식 승착은 흑160 절단이었습니다. 열 수 뒤의 흑180 빈삼각은 신진서가 승리를 감지한 수였습니다. 빈삼각이 역전의 한 수였다는 뜻이 아니라, 우세한 형세에서 모양의 금기보다 판의 요구를 우선한 판단의 상징이라는 뜻입니다. 신진서는:

- 일반 원칙을 알고 있었고

- 현재 형세와 KataGo의 반격 가능성을 읽었고
- 일부 손해를 감수하면서 우세를 지켰고
- 모양이라는 휴리스틱보다 승리라는 상위 목적을 우선했습니다.

즉 빈삼각 한 수에는: **상태 인식**

- 확률 평가
- 위험 관리
- 목적함수
- 휴리스틱 파괴

가 모두 들어 있습니다. 그래서 그 수가 아름답습니다. 모양이 아름다워서가 아니라: **상황에 대한 판단구조가 아름답습니다. 이세돌의 68수도 이제 새롭게 보인다** 그 수의 가치는: 당장 점수를 많이 얻었다. 가 아닙니다. 이세돌의 사후 회고를 이 프로젝트의 언어로 해석하면, 상대의 미래 선택환경을 바꿔: **아직 드러나지 않은 취약성을 시험할 상황을 만들었다.** 는 데 있을 수 있습니다. 따라서 전략의 가장 높은 층은: **내가 좋은 수를 두는 것 이 아니라 상대가 어려운 결정을 해야 하는 세계를 만드는 것** 입니다. **이건 엄청난 이동이다** 처음 질문: 어떤 수를 둘까? 그다음: 어떤 전략을 쓸까? 더 깊게: 어떤 판을 만들까? 그리고 인간에게는: **어떤 나를 만들까?** 까지 갑니다. 이 네 층을 구별해야 합니다. **MASTER MODEL v0.4 LAYER A — MOVE** 지금 무엇을 선택할 것인가. **LAYER B — POLICY** 반복적으로 어떤 선택규칙을 사용할 것인가. **LAYER C — ENVIRONMENT** 어떤 선택이 쉽게 일어나는 세계를 만들 것인가. **LAYER D — SELF** 어떤 선택을 자연스럽게 하는 사람이 될 것인가. 이렇게 되면: **의사결정** → 전략 → 환경설계 → 자기형성 이라는 새로운 축이 생깁니다. **NEXT QUESTION** 이제 바로 다음 단계가 매우 중요합니다. 지금까지 계속: 내가 선택한다. 라고 말했습니다. 그런데 신경과학과 철학은 묻습니다. **정말 ‘내가’ 선택하는가?** 결정 직전의 뇌에서는 이미:

- 기억
- 습관
- 감정
- 보상예측
- 신체상태
- 사회적 영향

이 움직이고 있습니다. AI도 수많은 학습결과를 통해 다음 행동을 출력합니다. 그렇다면 인간의 선택은 정말 AI의 정책 출력과 본질적으로 다른가? 그리고 모든 것이 이전 상태와 물리법칙에서 나온다면: **자유의지는 어디에 들어가는가?** 다음 장은 결국 이 문제로 가야 합니다. **선택하는 ‘나’는 어디에 있는가** 바둑기사의 한 수에서 자유의지·결정론·뇌·AI 정책·자기통제까지 여기부터 **바둑의 알고리즘이 본격적으로 인간 존재론을 공격하기** 시작합니다.

제2부

목표를 바꾸는 존재

결과를 고르는 능력에서 목표와 기준 자체를 검토하는 능력으로 이동한다.

3장. 자유와 책임

모든 행동에 원인이 있다면 선택과 책임은 왜 남는가?

자유는 원인에서 벗어난 빈 공간이 아니라 이유를 이해하고 충동을 조절하며 자신의 행동을 수정할 수 있는 행위자 구조에서 찾을 수 있다. 선택하는 '나'는 어디에 있는가 자유의지 · 결정론 · 뇌 · 의식 · 자기통제 · AI 정책 **TARGET CLAIM:** 뇌가 의식보다 먼저 움직인다면 자유의지는 환상이다 **신진서가 돌을 집어 든다** 판을 오래 봅니다. 손이 움직입니다. 돌이 내려갑니다. 우리는 자연스럽게 말합니다. 신진서가 이 수를 선택했다. 그런데 신경과학자가 옆에서 묻습니다. **정확히 언제 선택했습니까?** 손이 움직인 순간? “여기다”라고 느낀 순간? 그 생각이 의식에 올라오기 전? 몇 초 전부터 진행된 뇌 활동? 수십 년 동안 쌓인 바둑 훈련? 태어나면서 가진 성향? 질문을 이렇게 밀어붙이면: **'선택하는 나'의 위치가 갑자기 사라진다. 여기서 대중적인 자유의지 사망선고가 나왔다** 아주 매력적인 이야기입니다. 뇌가 먼저 결정 ↓ 의식은 뒤늦게 인식 ↓ 의식은 원인이 아님 ↓ **자유의지는 환상** 하지만 이 네 단계에는 거대한 논리적 도약이 있습니다. 특히: **준비전위: 이미 결정 완료** 라는 가정이 필요합니다. 그런데 바로 이 부분이 이후 연구에서 크게 흔들렸습니다. 2024년 대규모 리뷰는 리벳 이후 2,000편이 넘는 관련 문헌을 검토하면서, 준비전위의 의미, 의도 시점의 주관적 측정, 실험과 현실적 의사결정의 차이 등 여러 핵심 쟁점이 여전히 논쟁적이라고 정리합니다. **준비전위는 '결정 신호'가 아닐 수도 있다** 중요한 대안이 등장했습니다. 자발적 움직임 전에 뇌에는 여러 작은 변동이 계속 있습니다. 이 변동이 어떤 임계점에 도달하면: 지금 움직인다. 가 발생할 수 있습니다. 그러면 수백 번의 시행을: 실제 움직임 시점에 맞춰 평균내면, 움직임 이전에 점점 올라가는 것 같은 신호가 나타날 수 있습니다. 즉 우리가 본 준비전위가: 뇌가 오래전에 움직이기로 결정한 흔적 이라기보다 **여러 확률적 신경변동을 사건 시점에 맞춰 평균낸 결과** 일 수 있다는 것입니다.

실제로 준비전위를 사전 결정의 직접 증거로 보는 고전적 해석에는 여러 대안이 제시되어 왔고, 최근 리뷰들도 그 의미가 확정되지 않았다고 평가합니다. **사실 '뇌가 먼저'라는 표현부터 이상하다** 우리는 이렇게 말합니다. 내가 결정하기 전에 뇌가 결정했다. 그런데: **나는 뇌와 별개의 무엇인가?** 이 문장에는 은근히 작은 이원론이 숨어 있습니다. 마치: 나 와 내 뇌 가 서로 경쟁하는 두 존재인 것처럼 말합니다. 하지만 물리주의적 관점에서 인간의 판단은: **뇌·몸·기억·환경이 함께 만드는 과정** 입니다. 그러므로 뇌활동이 의식적 경험보다 먼저 시작됐다는 것은: 내가 아닌 무언가가 대신 결정했다. 는 뜻이 아닐 수 있습니다. 오히려: **나의 결정 과정이 의식에 올라오기 전에 이미 시작되어 있었다** 는 뜻일 수 있습니다. 여기서 **'나'의 정의가 중요해진다** 만약 나를: 지금 의식적으로 떠오르는 생각 으로만 정의한다면 무의식적 뇌처리는: **내가 아닌 것** 이 됩니다. 하지만 나를:

- 기억
- 습관
- 성격
- 감정
- 가치

- 신체
- 학습
- 의식

이 지속적으로 결합된 시스템으로 본다면? 그러면 무의식적인 준비도: **나의 일부**입니다. 여기서 **자유意志의 첫 번째 정의가 무너진다** 아주 강한 자유의지 정의: 어떤 원인에도 영향을 받지 않고 순수한 자아가 결정을 만든다. 이런 자유는 신경과학 이전부터 철학적으로도 이해하기 어렵습니다. 왜냐하면 원인이 전혀 없다면: **그것은 자유보다 무작위에 가까울 수 있기 때문입니다.** 내 성격도, 이유도, 가치도, 기억도 선택에 영향을 주지 않는다면? 왜 그 선택이: **나의 선택**입니까? **원인이 없다고 자유로운 것은 아니다** 이것은 자유의지 논쟁의 핵심 역설입니다. A: 모든 행동은 원인이 있다. 그러면 결정론처럼 보입니다. B: 어떤 행동에는 원인이 없다. 그러면 우연입니다. 그렇다면: **결정론도 자유가 아니고 무작위도 자유가 아니라면 자유는 어디에 있는가?** 여기서 고전적인 자유의지 문제가 본격적으로 등장합니다. **결정론부터 정확히 구분하자** 결정론은 대략: 특정 시점의 세계 상태와 자연법칙이 완전히 주어지면 이후 상태도 정해진다. 는 주장입니다. 만약 이것이 인간에게도 그대로 적용된다면: 유전자 + 환경 + 과거 경험 + 현재 뇌 상태 가 주어졌을 때 다음 선택도 결정될 것처럼 보입니다. 그러면 질문: 나는 정말 다르게 할 수 있었는가? 가 생깁니다. 그런데 **양자적 무작위성을 넣어도 문제가 해결되지 않는다** 어떤 사람이 말합니다. 양자역학은 확률적이니까 자유의지가 가능하다. 하지만: 내 뉴런이 양자적 우연 때문에 A 대신 B를 선택했다고 합시다. 그게 왜: **내가 자유롭게 B를 골랐다** 는 뜻입니까? 우연은: 통제 가 아닙니다. 따라서: **INDETERMINISM ≠ FREE WILL** 입니다. 그러면 **자유란 원인의 부재가 아니라 원인의 구조일 수 있다** 자유로운 행위는:

- 외부강제가 적고
- 자신의 이유를 고려하고
- 장기적 목표와 비교하고
- 충동을 억제할 수 있고
- 필요하면 행동규칙을 수정하는

시스템에서 나옵니다. 즉: **FREE WILL → SELF-GOVERNANCE** 로 이동합니다. '원인 없는 선택'보다: **자기조절**에 가까워집니다. 이때 **자기통제가 중요해진다** 어떤 사람이 말합니다. 케이크 먹고 싶다. 욕망이 있습니다. 동시에: 건강 때문에 먹지 않겠다. 라는 장기목표도 있습니다. 여기에는 서로 다른 수준의 선호가 충돌합니다. 즉: 즉각적 욕구 vs 장기적 자기정책 입니다. 최근 자기통제 연구도 자기통제를 단순한 하나의 '의지력 연료'라기보다 전략, 믿음, 목표 유지와 관련된 복합적인 과정으로 다룹니다. **예를 들어 담배 한 개비** 순간적으로: 피울까 말까? 만 보면 충동이 강합니다. 하지만 사람은: **POLICY** "나는 담배를 안 피운다." 라고 규칙을 만들 수 있습니다. **ENVIRONMENT** 집에 담배를 두지 않습니다. **SELF** "나는 비흡연자다." 라는 자기정체성을 구축합니다. 그러면 미래의 순간적 선택들이 달라집니다. 즉 인간은: **현재 행동만 선택하는 것이 아니라 미래의 선택을 만드는 시스템을 선택할 수 있습니다.** 이것은 엄청나게 중요한 자유의 형태다 자유를: 매 순간 아무거나 할 수 있음 으로 이해하면 약합니다. 더 강한 자유는: **미래의 내가 무엇을 하기 쉬운 사람인지 설계하는 능력** 일 수 있습니다. 이것은:

- 습관
- 교육

- 환경설계
- 자기규칙
- 약속
- 제도

로 구현됩니다. **자율성에는 시간축이 있다** 지금: 마음대로 먹는다. 는 자유와 10년 뒤: 건강해서 원하는 것을 할 수 있다. 는 자유가 충돌할 수 있습니다. 따라서: **자유(t0) ≠ 자유(t1)** 입니다. 일부 자기제약은: **미래의 자유를 사는 비용** 일 수 있습니다. **이것은 음악에서도 같다** 피아니스트가: 나는 자유롭게 연주하고 싶으니 음계를 연습하지 않겠다. 고 하면 자유가 커질까요? 아닙니다. 기초기술을 내면화할수록: **표현 가능한 선택공간이 커집니다**. 즉: discipline → capacity → freedom 경로가 있습니다. **여기서 자유를 다시 정의할 수 있다 자유 ≠ 선택지 숫자** 입니다. 선택지가 1,000개 있어도:

- 이해하지 못하고
- 실행할 능력이 없고
- 충동을 통제하지 못하면

실질적 자유는 작습니다. 따라서 자유에는: **선택가능성 + 이해능력 + 실행능력 + 수정능력** 이 필요합니다. **그러나 인간의 정책은 여러 층에서 수정될 수 있다** 바둑 AI의 목표: 승리. 는 외부에서 주어집니다. 반면 인간은: 나는 이런 바둑을 계속 두고 싶은가? 를 물을 수 있습니다. 더 넓게: 프로기사를 계속 할 것인가? 승리보다 무엇을 중요하게 여길 것인가? 까지 묻습니다. 즉 인간에게는: **policy evaluation** 뿐 아니라 **policy-of-policy evaluation** 이 있습니다. **메타인지가 등장한다** 나는 화가 납니다. 동시에: 내가 지금 화가 나서 판단이 흐려지고 있구나. 라고 생각할 수 있습니다. 이것은 매우 특별합니다. 시스템 안에: **자기 자신의 상태를 모델링하는 또 하나의 층** 이 있습니다. 행동하고 싶은 나 를 관찰하는 나 가 있습니다. 물론 이것도 뇌의 과정입니다. 하지만 기능적으로는 중요한 차이를 만듭니다. **이것이 emergent agency라는 아이디어로 이어진다** 물 분자 하나에는: 파도가 없습니다. 하지만 수많은 물 분자가 상호작용하면: **파도** 라는 상위 현상이 생깁니다. 마찬가지로 개별 뉴런에는: 인생계획 이 없습니다. 하지만 신경계 전체에서는: **목표 자기통제 계획** 이 나타날 수 있습니다. 따라서: **상위 수준의 설명은 하위 수준으로 환원 가능하다고 해서 무의미해지지 않는다**. 이것은 **바둑판에서도 보인다** 돌 하나에는: 세력 이 없습니다. 하지만 돌들의 관계에서는: **세력** 이 생깁니다. 돌 하나에는: 두터움 도 없습니다. 전체 관계 속에서 나타납니다. 즉: 개별 돌 → 관계 → 형태 → 전략적 의미 입니다. **그러면 자유의지는 무엇으로 남는가** 현재까지 가장 강한 후보는: **반응능력(responsiveness to reasons)** 입니다. 즉 사람에게 이유를 제시했을 때:

- 이해하고
- 비교하고
- 행동을 수정하고
- 장기목표에 맞게 충동을 억제할 수 있다면

그 행위는 더 자유롭다고 볼 수 있습니다. 이런 방향은 철학의 **양립가능론(compatibilism)**과 잘 연결됩니다. 결정론이 참이더라도 특정 종류의 자기통제와 이유에 대한 반응성을 자유로 볼 수 있다는 입장입니다. 자유의지에 대한 현대 철학적 논쟁 역시 자유를 단순한 '원인 없는 선택'으로만 다루지 않습니다. **그러면 강제와 자유를 연속선으로 볼 수 있다** 낮은 자율성 층으로 협박당함. ↓ 중독적 충동.

↓ 강한 습관. ↓ 사회적 압력. ↓ 충분한 숙고. ↓ 자기 가치와 일치한 선택. ↓ 장기적 자기형성까지 포함한 선택. 높은 자율성 즉 자유는: ON / OFF 보다 DEGREE 문제일 수 있습니다. 교육의 목적도 달라진다 교육을: 지식을 많이 넣는 것 이라고 보면 좁습니다. 더 깊게는: 더 많은 좋은 선택을 할 수 있는 행위자를 만드는 것 입니다. 교육은:

- 세계모델을 넓히고
- 선택지를 늘리고
- 결과를 예측하게 하고
- 거짓 정보를 구별하게 하고
- 자기 욕망을 반성하게 합니다.

즉: **education** → **agency expansion** 입니다. 돈도 다시 보인다 돈이 자유를 사는 이유는: 원하는 물건을 많이 살 수 있어서 만이 아닙니다. 돈은: 선택공간을 넓힙니다.

- 직장을 그만둘 수 있음
- 이동할 수 있음
- 기다릴 수 있음
- 치료를 받을 수 있음
- 새로운 실험을 할 수 있음

그러므로 앞 장의: **option space** 와 다시 연결됩니다. 이제 AI 시대의 자유 문제가 나타난다 과거에는 주로: AI가 우리 대신 결정을 내릴까? 를 걱정했습니다. 더 미묘한 문제는: AI가 우리가 무엇을 선택할 수 있다고 느끼는지 설계하는 것 입니다. 추천, 검색순위, 뉴스피드, 가격, 개인화, 자동완성. 이런 시스템은: 선택 전에 후보군을 만든다. 그러므로 AI 시대의 자유는 '최종 클릭'에만 있지 않다 최종적으로 사람이 버튼을 눌렀다고: 인간이 자유롭게 선택했다. 고 끝낼 수 없습니다. 앞에서:

- 후보군
- 순위
- 프레이밍
- 기본값
- 추천

이 이미 만들어졌을 수 있습니다. 따라서 자유를 분석하려면: **OPTION GENERATION** → **OPTION VISIBILITY** → **EVALUATION** → **CHOICE** → **REVISION** 전체를 봐야 합니다. 이것은 바둑 AI와 인간의 관계에서도 이미 일어났다 AI 이전 프로기사: 인간 경험 → 후보수 AI 이후 프로기사: 인간 경험 + AI 추천 → 후보수 입니다. 즉 AI가 단순히: 더 좋은 정답을 알려준다. 가 아닙니다. 인간이 생각할 수 있는 바둑 자체를 바꿉니다. 그러면 인간 자유는 AI 때문에 줄어드는가? 반드시 그렇지 않습니다. AI가:

- 새로운 후보를 보여주고
- 기존 편견을 깨고
- 결과를 더 잘 예측하게 하면

인간의 실질적 선택능력 이 커질 수 있습니다. 즉: AI → agency expansion 도 가능합니다. 동시에 반대도 가능하다 AI가:

- 후보를 지나치게 좁히고
- 인간이 평가하지 않고 따르게 하고
- 스스로 판단하는 능력을 약화시키면

agency contraction 도 가능합니다. 따라서 중요한 질문은: AI가 결정을 대신하느냐? 보다: AI를 사용한 뒤 인간의 판단능력이 커졌는가 작아졌는가? 입니다. 계산기 비유 계산기는 인간에게 긴 나눗셈을 대신합니다. 그래서 일상 계산은 쉬워졌습니다. 하지만: 계산기의 숫자가 말이 되는가? 를 판단할 수 없다면 의존적입니다. AI도 마찬가지입니다. 가장 좋은 관계는: AI가 계산능력을 확장하되 목적·검증·수정권은 인간에게 남기는 것 일 수 있습니다. MASTER MODEL v0.5 — AGENCY 자유로운 행위자를 현재까지 이렇게 정의해볼 수 있습니다. ① MODEL 세상에 대한 모델을 가진다. ② SELF-MODEL 자신의 욕망·한계·성향을 어느 정도 안다. ③ OPTION GENERATION 대안을 만들 수 있다. ④ REASONING 대안의 이유와 결과를 비교한다. ⑤ INHIBITION 즉각적 반응을 멈출 수 있다. ⑥ COMMITMENT 장기 목표에 따라 정책을 만들 수 있다. ⑦ ENVIRONMENT DESIGN 미래 선택환경을 조정할 수 있다. ⑧ META-REVISION 자신의 목표·정책·가치를 다시 검토할 수 있다. 이 능력들의 묶음을: AGENCY 라고 부르겠습니다. 그러면 자유는 신비한 한 순간에서 사라진다 리벳 실험은 물었습니다. 의식적으로 움직이겠다고 느낀 순간이 언제인가? 하지만 인간 자유가 앞 장의 구조라면: 200밀리초짜리 순간에 자유 전체가 들어있을 이유가 없다. 자유는:

- 어린 시절 학습
- 어제의 계획
- 한 시간 전 결심
- 현재의 억제
- 행동 이후의 반성

에 걸쳐 분산되어 있을 수 있습니다. 따라서 리벳형 실험이 자유의지 전체를 판정하기 어려운 이유도 선명해집니다. 현대 리뷰 역시 단순 손가락 움직임 과제가 복잡한 자발적 행위 전체를 대표할 수 있는지에 상당한 주의를 요구합니다. 그리고 이것은 인간 전체에도 적용될 수 있다 사람이 어떤 선택을 합니다. 우리는: 왜 그랬어? 라고 묻습니다. 하지만 그 행동의 원인은 한 순간에 있지 않을 수 있습니다. 그 사람의:

- 성장
- 사회
- 언어
- 기억
- 관계
- 몸
- 가치

가 모두 들어 있습니다. 그렇다고 책임이 자동으로 사라지는 것은 아닙니다. 오히려 새로운 질문이 생깁니다. 이 시스템은 자신의 행동을 이해하고 수정할 능력이 있었는가? 바로 여기서 ‘책임’ 문제가

열린다 만약 인간이 완전히 원인의 산물이라면: 범죄자를 왜 처벌하는가? 라는 질문이 나옵니다. 반대로 인간에게 무조건 자유의지가 있다고 선언하면:

- 중독
- 뇌손상
- 강압
- 발달
- 정신상태

의 차이를 설명하기 어렵습니다. 그러므로 책임 역시: **0 또는 1** 이 아니라 **행위능력의 정도** 와 연결될 가능성이 있습니다. 모든 행동에 원인이 있다면 책임은 왜 남는가 자유의지 · 비난 · 처벌 · 억지 · 교정 · 책임능력 · 사회적 피드백 **TARGET CLAIM:** 행동이 유전자·환경·뇌 상태·과거 경험의 결과라면 누구도 책임질 수 없다 **한 사람이 돌을 잘못 두었다** 바둑에서는 아주 자연스럽게 말합니다. 악수였다. 그리고 복기합니다. 왜 그 수를 두었는지 봅니다.

- 형세판단을 잘못했는가
- 수읽기를 놓쳤는가
- 시간에 쫓겼는가
- 상대의 의도를 오해했는가
- 평소의 잘못된 습관이 나왔는가

이상한 점이 있습니다. 우리는 그 악수에: **원인이 있다는 것을 찾으면서도 동시에: 그 기사에게 책임을 묻습니다.** “원인이 있었으니까 당신 잘못이 아니다.” 라고 하지 않습니다. 오히려 원인을 알아야: 다음에는 어떻게 고칠 것인가? 를 말할 수 있습니다. 이 장면이 책임의 본질에 관한 중요한 단서를 줍니다. **책임에는 서로 다른 질문들이 섞여 있다** 우리가: 네 책임이야. 라고 말할 때 사실 최소 네 가지를 말할 수 있습니다. ① *CAUSAL RESPONSIBILITY* 원인 책임 네 행동이 결과를 일으켰는가? ② *CAPACITY RESPONSIBILITY* 능력 책임 당시 무엇을 하는지 이해하고 행동을 조절할 능력이 있었는가? ③ *MORAL RESPONSIBILITY* 도덕적 책임 이 행동에 대해 비난받을 만한가? ④ *INSTITUTIONAL RESPONSIBILITY* 제도적 책임 사회가 어떤 제재·배상·교정·격리를 부과해야 하는가? 이 네 가지는 같습니다가 아닙니다. **반대로 인간에게 책임을 묻는 이유** 사람에게는 대체로:

- 규칙을 이해하는 능력
- 결과를 예상하는 능력
- 금지 이유를 이해하는 능력
- 충동을 억제하는 능력
- 처벌 가능성을 고려하는 능력
- 이후 행동을 수정하는 능력

이 있습니다. 바로 앞 장에서 만든: **AGENCY** 의 일부입니다. 현대 자유의지 철학에서도 책임과 연결되는 중요한 후보 가운데 하나가 **이유에 반응하는 통제 능력**입니다. 모든 원인이 없는 존재여야 책임이 있다는 식으로만 논의되지는 않습니다. **이것이 아주 중요한 전환이다** 오래된 책임관을 극단적으로 표현하면: 그는 다른 선택을 할 수 있었는데 악을 선택했다. 따라서: **고통을 받을 자격이 있다.** 입니다. 그런데 결정론을 강하게 받아들이면 이 구조는 흔들립니다. 그 사람의:

- 유전자
- 성장환경
- 충동성
- 교육
- 뇌 상태
- 과거 경험

까지 모두 원인의 사슬 안에 있다면: 완전히 독립적인 최초의 선택자가 어디 있느냐? 는 문제가 생깁니다. 실제로 자유의지와 도덕적 책임의 관계는 현대 철학에서도 계속 논쟁 중이며, 결정론과 자유의 양립 여부와 책임이 가능한지는 구별해서 다뤄야 한다는 점이 강조됩니다. **여기서 처벌의 목적을 분해해야 한다** 형벌철학과 실제 형사제도에는 여러 목적이 섞여 있습니다. 대표적으로: *RETRIBUTION* 응보 잘못했으므로 그에 상응하는 벌을 받아야 한다. *DETERRENCE* 억지 처벌 가능성이 앞으로 범죄를 줄인다. *INCAPACITATION* 격리 위험한 사람이 추가 피해를 만들 수 없도록 제한한다. *REHABILITATION* 교정·재활 행동능력과 환경을 바꿔 재범 가능성을 낮춘다. 미국 연방 형사법 맥락에서도 이 네 가지가 전통적인 주요 양형 목적들로 명시적으로 구분됩니다. 중요한 것은: **이 네 가지의 논리가 서로 다르다. 특히 응보는 독특하다** 억지는 미래를 봅니다. 격리도 미래를 봅니다. 교정도 미래를 봅니다. 하지만 응보는 말합니다. 미래에 도움이 되느냐와 별개로 과거의 잘못 때문에 벌을 받을 자격이 있다. 즉: **DESERT** 문제입니다. 바로 여기에서 자유의지 문제가 가장 날카롭게 들어옵니다. **뇌손상 사례가 이것을 아주 분명하게 보여준다** 어떤 사람에게 전두엽 손상이 생긴 뒤:

- 충동조절
- 판단
- 사회적 행동

이 크게 변했다고 합시다. 행동 결과만 보면 똑같은 폭행입니다. 하지만 우리는: 건강했던 사람이 계획적으로 폭행한 경우 와 똑같이 보지 않습니다. 왜냐하면: **행위능력이 다를 수 있기 때문입니다.** 신경학적 증거가 형사책임, 재판능력, 감경 문제에서 실제로 점점 활용되고 있다는 점도 바로 이 경계의 중요성을 보여줍니다. **따라서 책임에는 정도가 있을 수 있다** 앞 장에서 자유를: 0 / 1 이 아니라 정도 문제로 보았습니다. 책임도 마찬가지로 가능성이 큼니다. 예를 들어: **매우 낮은 책임능력**

- 심각한 의식장애
- 강한 강제
- 행동의 의미를 이해할 능력이 현저히 손상

↓ 부분 책임

- 상당한 충동조절 장애
- 발달적 미성숙
- 일부 판단능력 손상

↓ 높은 책임

- 충분한 이해

- 계획성
- 대안 인식
- 자기통제 가능
- 의도적 실행

입니다. 즉: **RESPONSIBILITY $\approx$ CAPACITY-SENSITIVE** 입니다. 이제 처벌을 '피드백'으로 바라보자 여기서 바둑으로 돌아갑시다. 기사가 악수를 두었습니다. 다음 날 감독이: 네가 악수를 두었으므로 손가락을 때리겠다. 고 한다면 이상합니다. 더 자연스러운 대응은: 복기 입니다. 행동 → 결과 → 원인분석 → 정책수정 입니다. 이것이 학습입니다. **사회도 어느 정도 같은 구조를 가진다** 사람이 규칙을 어깁니다. 사회가:

- 비난
- 벌금
- 손해배상
- 자유 제한
- 교육
- 치료

등으로 반응합니다. 그 반응은 다시:

- 당사자의 행동
- 주변 사람의 기대
- 사회적 규범

에 영향을 줍니다. 즉: **ACTION → CONSEQUENCE → SOCIAL FEEDBACK → BEHAVIOR UPDATE** 입니다. **경제학 언어로 보면 아주 명확하다** 행동 A의 기대효용이: +10 이었다고 합시다. 사회가 그 행동에: -20 의 예상 제재를 붙입니다. 그러면: $+10 - 20 = -10$ 이 되어 행동 유인이 바뀔 수 있습니다. 즉 법은 단순히: **나쁜 행동 이후에 등장** 하는 것이 아닙니다. 행동 이전의: **선택구조** 를 바꿉니다. **하지만 여기서 중요한 한계** 사람이 모든 행동을 이런 계산으로 하는 것은 아닙니다. 범죄는:

- 충동
- 감정
- 중독
- 사회규범
- 도덕
- 기회
- 낮은 적발 가능성

등에도 영향을 받습니다. 따라서: **형량 $\uparrow$ → 범죄 $\downarrow$ 가 자동 법칙은 아닙니다.** 사회적 피드백은: **인간이 실제로 반응할 수 있는 방식** 과 맞아야 합니다. **교정은 더 직접적이다** 교정은: 다시는 하지 마. 에서 끝나지 않습니다. 행동을 만든 메커니즘을 바꾸려 합니다. 예: 충동조절 부족 → 치료·훈련 약물의존 → 중독치료 기술부족 → 직업교육 범죄 네트워크 → 환경 분리 즉: **행동 출력** 보다 **행동 생성 시스템** 을 수정합니다. **바로 앞 장의 POLICY / ENVIRONMENT / SELF** 다 범죄가 한 번의 MOVE였다면:

사회는 단순히 그 MOVE를 벌하는 것만으로 충분하지 않을 수 있습니다. *POLICY* 행동규칙 자체를 바꾼다. *ENVIRONMENT* 범죄가 일어나기 쉬운 조건을 바꾼다. *SELF* 충동조절·능력·정체성을 바꾼다. 그러면 형사정책은: **MOVE PUNISHMENT** 에서 **POLICY / ENVIRONMENT REPAIR** 까지 확장됩니다. 그래서 **책임은 여전히 필요하다** 책임은 단순한 형벌철학이 아닙니다. 책임을 묻는다는 것은 사람에게: 당신은 이유를 이해하고 답할 수 있는 존재다. 라고 대하는 일이기도 합니다. 즉 책임은 역설적으로: **행위자를 인간으로 인정하는 것** 일 수도 있습니다. 모든 잘못을: 뇌가 그렇게 했어요. 라고 해버리면 사람을 책임에서 해방시키는 동시에: **행위자로서의 지위까지 빼앗을 위험** 이 있습니다. 이것이 **책임의 양면성**이다 책임은: 부담 비난받고 배상하고 제재받습니다. 동시에: **지위** 내 행동을: **내 이유와 선택의 결과** 로 취급받습니다. 따라서 완전히 책임 없는 존재로 취급되는 것은 반드시 존중이 아닙니다. 아이, 중증 인지장애, 기계 에 대해서는 일반 성인과 다른 방식으로 책임을 묻습니다. 그러므로 **책임을 전부 공학으로 환원할 수 없다** “어떤 벌이 재범률을 가장 많이 낮추나?” 만 보면: **사회공학** 문제가 됩니다. 하지만 피해자는 묻습니다. 나에게 이런 일을 해놓고 아무 책임도 없는가? 즉: **공정성** 과

인정 문제가 있습니다. 따라서 형사제도를 완전히: 행동교정 알고리즘 으로 환원하면 중요한 인간적 차원을 놓칩니다. 현재 가장 강한 구조 책임제도에는 최소 네 층이 필요합니다. ① **ACCOUNT** 무슨 일이 있었는가. ② **CAPACITY** 행위자가 무엇을 이해하고 통제할 수 있었는가. ③ **ANSWERABILITY** 그 이유에 대해 설명하고 응답할 수 있는가. ④ **CONSEQUENCE** 피해회복·제재·보호·교정을 어떻게 할 것인가. 이렇게 보면: **책임 ≠ 단순한 비난** 입니다. 여기서 ‘**후회**’도 다시 보이다 인간의 후회는 내부 처벌처럼 작동할 때가 있습니다. 그때 왜 그랬을까. 그러나 좋은 후회는: **자기를 괴롭히는 것** 에서 끝나지 않습니다. 무엇을 놓쳤는가 → 다음에는 무엇을 바꿀 것인가 로 갑니다. 즉: **REGRET** → **UPDATE** 입니다. **나쁜 후회는 업데이트가 없다** 계속: 나는 왜 이 모양이지. 라고만 합니다. 그러면: 오류 → 자기비난 → 고통 에서 멈춥니다. 좋은 복기는: 오류 → 메커니즘 발견 → 정책 수정 입니다. 따라서: **책임을 목적에는 학습이 포함될 수 있다. 기업도 똑같다** 제품 사고가 났습니다. 나쁜 조직: 누가 잘못했어? 한 명을 찾아 처벌합니다. 끝. 좋은 조직: 누가 잘못했는가? 도 보지만 동시에:

- 왜 이 오류가 발견되지 않았는가
- 왜 보고되지 않았는가
- 왜 시스템이 한 사람 실수에 취약했는가

를 봅니다. 즉: **PERSON ERROR** 에서 **SYSTEM ERROR** 까지 올라갑니다. **항공사고 조사가 좋은 비교가 된다** 조종사가 실수했을 수 있습니다. 그러나 조사에서는:

- 인터페이스
- 훈련
- 피로
- 절차
- 조직문화
- 기계설계

까지 봅니다. 왜냐하면 목적이: **다음 사고 방지** 이기 때문입니다. 즉: **누구의 잘못인가?** 를 넘어서 **어떻게 이런 결과가 가능했는가?** 를 묻습니다. 여기서 **사회의 지능**이라는 개념이 나타난다 지능적인

개인은: 오류 → 학습 합니다. 지능적인 조직도: 사고 → 조사 → 제도 변경 합니다. 지능적인 사회도: 문제 → 원인 분석 → 법·제도 수정 합니다. 그러면 새로운 가설이 생깁니다. **문명의 지능은 오류를 얼마나 잘 피드백으로 변환하는가에 달려 있을 수 있다.** 여기서 책임과 학습이 충돌할 수도 있다 누군가 실수를 보고하면: 무조건 처벌 하는 조직을 생각해봅시다. 사람들은: **실수를 숨깁니다.** 그러면 책임을 강하게 묻는 시스템이 오히려: **학습능력을 떨어뜨릴 수 있습니다.** 이것이 중요한 역설입니다. **반대로 아무 책임도 없으면?** 괜찮아, 실수할 수 있지. 만 반복합니다. 그러면:

- 부주의
- 반복 위반
- 위험 행동

에 비용이 없습니다. 역시 학습이 약해질 수 있습니다. 따라서 좋은 제도는: **책임과 학습을 동시에 살려야 한다.** AI 책임 문제에 들어가면 더 재미있어진다 자율주행차가 사고를 냈습니다. 누구 책임입니까?

- 승객?
- 운전자?
- 자동차 회사?
- 모델 개발자?
- 데이터 공급자?
- 시스템 통합자?

AI에는 인간처럼: **도덕적 비난** 을 바로 적용하기 어렵습니다. 하지만: **인과적 책임** 은 추적할 수 있습니다. 이것이 AI 시대 책임의 핵심 문제다 전통적 책임 모델: 한 사람 → 한 행동 → 한 결과 AI 시스템에서는: 데이터 + 모델 + 개발자 + 배포기업 + 사용자 + 제도 ↓ 결과가 됩니다. 즉: **RESPONSIBILITY DISTRIBUTION** 문제가 생깁니다. 따라서 책임은 인터페이스를 따라 추적해야 한다 누가:

- 목표를 정했는가
- 데이터를 선택했는가
- 모델을 배포했는가
- 위험한 출력을 차단할 권한이 있었는가
- 최종 결정을 수정할 능력이 있었는가

를 봐야 합니다. 즉 책임은: **CONTROL + KNOWLEDGE + CAPACITY TO INTERVENE** 와 연결됩니다. **아주 강한 원칙 후보가 나온다 책임은 영향력과 통제가능성을 따라가야 한다.** 아무 통제권이 없는 사람에게 모든 책임을 주는 것은 부당합니다. 반대로 막대한 통제권을 가진 주체가: 기계가 결정했으므로 책임 없다. 라고 하는 것도 이상합니다. 이제 사회를 하나의 학습 시스템으로 볼 수 있다 사회의 구조를 아주 추상화해봅시다. 행동 ↓ 결과 ↓ 피해·보상 ↓ 평가 ↓ 책임 귀속 ↓ 피드백 ↓ 행동규칙 수정 ↓ 제도 수정 ↓ 다음 행동 이 됩니다. 이것은 머신러닝의 학습루프와 형태적으로 닮았습니다. 하지만 사회에는 결정적 차이가 있다 AI의 손실함수는 설계자가 정합니다. 사회에서는: 무엇이 손실인가를 사람들이 다룹니다.

- 얼마나 처벌할 것인가

- 누구의 자유를 보호할 것인가
- 피해를 어떻게 평가할 것인가
- 평등과 효율 중 무엇을 우선할 것인가

따라서 사회적 학습에는: **정치와 윤리** 가 들어갑니다. **오히려 사회의 핵심은 '누가 보상함수를 정하는가'다** 왕정에서는: 왕 의 가치가 강할 수 있습니다. 독재에서는: 지배집단. 민주주의에서는: 경쟁하는 가치들이 제도적으로 협상됩니다. 즉 정치란 상당 부분: **COLLECTIVE VALUE FUNCTION FORMATION** 문제라고 볼 수 있습니다. 이것은 단순한 비유 이상으로 상당히 강력합니다. **바둑과 다시 비교하면 엄청난 차이가 드러난다** 바둑: 승리한다. 모두가 목적함수에 합의합니다. 사회: 좋은 사회란 무엇인가? 에서부터 싸웁니다. 성장? 평등? 자유? 안전? 전통? 혁신? 그러므로: **사회는 최선의 수를 찾기 전에 무엇을 승리라고 부를지를 결정해야 한다. 여기서 아주 중요한 문장이 살아남는다는 책임은 원인이 없는 사람을 찾아내는 일이 아니다.** 더 강하게 말하면: **책임은 원인의 사슬 속에서 어느 행위자가 이유를 이해하고 행동을 통제·수정할 능력과 권한을 가졌는지를 찾아, 결과를 미래의 더 나은 선택으로 되먹임하는 제도다.** 다만 이것만으로 응보적 '마땅함'의 문제까지 해결되는 것은 아닙니다. 그 부분은 철학적으로 계속 열려 있습니다. **그리고 이제 훨씬 더 큰 질문이 열린다** 한 사람이 잘못하면: 책임을 묻고 수정한다. 그렇다면 사회 전체가 잘못하면?

- 금융위기
- 전쟁
- 환경파괴
- 산업재해
- 정책실패

가 일어났을 때: **사회는 어떻게 복기하는가?** 그리고 더 무서운 질문: 어떤 사회는 왜 같은 실수를 반복하고, 어떤 사회는 한 번의 실패에서 제도를 바꾸는가? 여기서 바둑의 **복기**가 문명 전체로 확장됩니다.

4장. 학습과 되먹임

결과가 되 돌아온다는 사실만으로 개인과 문명은 학습하는가?

피드백은 결과의 귀환일 뿐이다. 학습은 차이를 감지하고 원인을 판별하며 모델·정책·규칙을 바꿔 다음 행동이 달라질 때 성립한다. 좋습니다. 이번 수정의 핵심은 네 층이 기능적으로 서로 침범하지 않게 경계를 잡는 것입니다. 특히 오류가 발생함과 오류가 보임, 오류가 보임과 원인을 앎, 원인을 앎과 실제로 제도를 바꿈을 분리하겠습니다. 4층 경계 강화 · 시장/민주주의/법 사례 비교 · 오류→수정 전달경로 CORE MODEL: 탐색 → 감지 → 판별 → 수정 단, 네 층은 단순한 순서가 아니라 서로 다른 기능을 담당하는 독립 단계다. 먼저 네 층의 경계를 한 문장씩 잠근다 LAYER 1 — 탐색 무엇을 시험할 것인가 새로운:

- 행동
- 정책
- 상품
- 주장
- 판결논리

를 현실에 내놓는다. 핵심 질문: 어떤 대안들이 실제로 시험되고 있는가? LAYER 2 — 감지 무엇이 예상과 달랐는가 시험된 행동의 결과를 관찰하고: 기대와 실제 사이에 차이가 발생했는가? 를 발견한다. 핵심 질문: 무언가 잘못되었다는 신호가 존재하는가? LAYER 3 — 판별 왜 달랐는가 관찰된 차이가:

- 우연인지
- 정책 때문인지
- 외부충격 때문인지
- 다른 메커니즘 때문인지

를 구별한다. 핵심 질문: 무엇을 수정해야 한다고 결론 내릴 근거가 있는가? LAYER 4 — 수정 그래서 무엇을 실제로 바꿀 것인가 살아남은 설명을:

- 행동
- 가격
- 정책
- 법
- 절차
- 조직
- 기억

에 반영한다. 핵심 질문: 다음번 시스템의 행동이 실제로 달라졌는가? 네 층을 혼동하면 무엇이 잘못되는가 가장 중요한 HARD LOCK은 이것이다. 대안 존재 ≠ 탐색 완료 아이디어가 존재해도

시험되지 않았다면 탐색이 아니다. **결과 발생** ≠ **감지** 문제가 생겼어도 누구도 보지 못했다면 감지가 아니다. **감지** ≠ **판별** 매출이 떨어졌다고 왜 떨어졌는지 아는 것은 아니다. **판별** ≠ **수정** 원인을 정확히 알아도 조직이 아무것도 바꾸지 않으면 학습이 아니다. 따라서: **사건** → **신호** → **설명** → **변화** 를 각각 분리해야 한다. 더 엄격한 네 층 정의 **LAYER 1 — 탐색** *Variation + Exposure* 탐색에는 두 조건이 동시에 필요하다. A. 차이가 존재해야 한다 서로 다른:

- 상품
- 정책
- 이론
- 후보
- 판결논리

가 있어야 한다. B. 현실에 노출되어야 한다 머릿속 아이디어만으로는 부족하다. 실제로:

- 판매되고
- 시행되고
- 투표되고
- 주장되고
- 적용되어야 한다.

따라서: **VARIATION WITHOUT EXPOSURE** ≠ **EXPLORATION** 이다. **감지층의 절대 금지선** 감지 단계에서: 왜 그랬는지 를 미리 넣으면 안 된다. 예: 매출 감소 → 가격이 비싸서 실패. 이것은 아직 판별이다. 감지가 확실히 말할 수 있는 것은: 예상한 만큼 팔리지 않았다. 까지다. 이 경계를 유지해야 확증편향을 막을 수 있다. **LAYER 3 — 판별** *Signal* → *Cause / Model* 판별은 가장 어려운 층이다. 하나의 신호에는 여러 설명이 가능하다. 매출 감소:

- 가격
- 품질
- 경쟁자
- 경기침체
- 유통
- 계절성

선거패배:

- 경제
- 후보
- 부패
- 외교
- 세대구조

판결 뒤집힘:

- 사실인정 오류

- 법해석 오류
- 절차오류
- 새로운 증거

따라서 판별은: **설명 경쟁** 이다. **판별이 진짜 판별이 되기 위한 조건** 최소한 다음이 필요하다. **경쟁가설** 다른 설명이 존재해야 한다. 증거 설명과 현실을 비교해야 한다. 반증 가능성 틀릴 수 없는 설명은 판별에 약하다. **인과구분** 동시발생과 원인을 구별해야 한다. 즉: **Signal $\neq$ Cause** 이다. **LAYER 4 — 수정** *Diagnosis $\rightarrow$ Changed System* 수정은: 교훈을 얻었다. 가 아니다. 실제로:

- 상품을 없애거나
- 가격을 바꾸거나
- 정부가 정책을 철회하거나
- 판례가 바뀌거나
- 법률이 개정되는

변화가 필요하다. 그리고 중요한 것이 하나 더 있다. **수정된 규칙이 저장되어야 한다.** 그렇지 않으면 몇 년 뒤 같은 실수가 반복된다. 따라서 수정층은: **UPDATE + RETENTION** 이다. 따라서 오류에서 수정까지는 자동으로 연결되지 않는다 핵심은: **ERROR $\neq$ SIGNAL $\neq$ DIAGNOSIS $\neq$ CORRECTION** 이다. 중간에 전달경로가 있다. 이를 가장 짧게 압축하면: **오류 발생 $\rightarrow$ 오류 가시화 $\rightarrow$ 오류 전달 $\rightarrow$ 오류 해석 $\rightarrow$ 수정 권한 $\rightarrow$ 실제 수정** 이다. 이 중 하나라도 끊기면 학습이 멈춘다. **시장 사례 01** 신제품이 팔리지 않는다 기업이 새로운 스마트폰을 출시했다고 하자. 예상: 첫 분기 100만 대. 실제: 40만 대. **LAYER 1 — 탐색** 기업이 기존 제품과 다른:

- 가격
- 디자인
- 기능

의 신제품을 출시했다. 즉: **새로운 상품가설을 현실에 노출** 했다. **LAYER 2 — 감지** 판매량: 예상 100만 vs 실제 40만 차이: **-60만** 여기까지가 감지다. 아직: 왜 안 팔렸는가? 는 모른다. **LAYER 3 — 판별** 가능한 설명: A. 가격이 비쌌다. B. 카메라 성능이 부족했다. C. 경쟁사 신제품 때문이었다. D. 경기침체 때문이었다. 기업은:

- 지역별 판매
- 경쟁제품
- 소비자조사
- 가격민감도

를 비교한다. 여기서 원인에 대한 경쟁이 일어난다. **LAYER 4 — 수정 판별** 결과: 제품 자체보다 가격이 핵심이었다. 라고 판단했다면:

- 가격 인하
- 보급형 출시
- 생산량 축소

등을 실행한다. 다음 분기 판매량을 다시 본다. 이때 비로소: **학습루프 완성** 이다. **시장에서 전달경로가 끊기는 사례** 현장 판매직원은: 고객들이 너무 비싸다고 한다. 를 알고 있다. 그러나 본사는: 브랜드 문제일 리 없다. 며 보고를 무시한다. 그러면: **ERROR EXISTS** 하지만: **ERROR DOES NOT REACH AUTHORITY** 이다. 즉 시장 자체가 신호를 만들었어도 기업 내부의 전달구조 때문에 학습이 실패할 수 있다. **민주주의 LAYER 2 — 감지** 다양한 센서가 작동한다.

- 주택가격 통계
- 전월세 통계
- 거래량
- 여론조사
- 언론
- 민원

여기에서: 목표했던 안정이 충분히 발생하지 않았다. 는 신호가 생긴다. 중요: **지지율 하락 자체는 정책 실패의 정확한 측정값이 아니다**. 왜냐하면 지지율은 여러 문제를 동시에 압축하기 때문이다. **민주주의 LAYER 3 — 판별** 정부: 세계적 유동성 때문이다. 야당: 정책 때문이다. 전문가 A: 공급부족 때문이다. 전문가 B: 금융조건 때문이다. 즉: **정치적 설명경쟁** 이 벌어진다. 여기서:

- 통계
- 지역비교
- 시간차
- 해외비교

등이 필요하다. 하지만 정치에서는 과학실험처럼 완벽한 대조군을 만들기 어렵다. 따라서 판별은 시장보다 훨씬 어렵다. **민주주의 LAYER 4 — 수정** 가능한 수정경로: 정책 수정 정부가 규제를 변경. 인적 수정 장관 교체. 권력 수정 선거를 통한 정권교체. 즉 민주주의에서는 수정의 단위가: **POLICY** 뿐 아니라 **GOVERNING AGENT** 까지 포함한다. 이것이 시장과의 중요한 차이다. **민주주의의 대표적 학습 실패** 국민 다수가 문제를 느낀다. 그러나:

- 언론이 왜곡하거나
- 정당이 의제를 흡수하지 못하고
- 선거제가 선택지를 막거나
- 권력교체가 어려우면

오류신호가 있어도: **수정으로 전달되지 않는다**. 따라서 민주주의에서 중요한 것은 단순 여론이 아니다. **여론 → 권력수정 인터페이스** 다. **법 사례 01 1심 판결이 뒤집힌다** 1심에서 피고인에게 유죄가 선고됐다. 피고인은 주장한다. 증거평가가 잘못됐다. 항소한다. **LAYER 1 — 탐색** 여기서 탐색은 시장처럼 상품을 여러 개 출시하는 것이 아니다. 법에서는:

- 서로 다른 법리
- 서로 다른 사실해석
- 원고와 피고의 경쟁 주장

이 제출된다. 즉: **competing legal interpretations** 가 시험된다. **법 LAYER 2 — 감지 오류** 신호는: 패소했다. 그 자체가 아니다. 중요한 감지채널은:

- 항소이유
- 새로운 증거
- 절차 위반 주장
- 판결이유의 모순

이다. 즉 당사자가: 이 판단에는 문제가 있다. 고 오류신호를 공식적으로 만든다. **법 LAYER 3 — 판별** 상급심:

- 증거
- 법률
- 절차
- 판결이유

를 검토한다. 여기서 묻는다. 사실을 잘못 인정했는가? 법을 잘못 해석했는가? 절차를 위반했는가? 즉 법에서는 판별이 매우 형식화되어 있다. **법의 전달경로** 당사자의 문제 경험 ↓ 소송·항소 ↓ 증거와 법률언어로 변환 ↓ 재판부 심리 ↓ 판결 이유 ↓ 상급심 검토 ↓ 판결 수정 ↓ 판례·법 개정 이다. 여기서 중요한 것은: **법에서는 오류가 법적으로 번역되어야 전달된다.** 단순히: 억울하다. 고 느끼는 것만으로는 충분하지 않다. **세 제도의 '센서'가 다르다** **MARKET SENSOR** 가격·구매행동 사람이 반드시 이유를 설명할 필요가 없다. 사지 않으면 된다. **DEMOCRATIC SENSOR** 의견·표현·투표 시민의 선호가 정치적 신호로 변환된다. **LEGAL SENSOR** 주장·증거·절차 오류를 법적 언어로 표현해야 한다. **전달경로 역시 다르다** 시장 소비자 → 구매/비구매 → 매출·가격 → 기업 → 제품·가격 수정 민주주의 시민 → 여론·언론·정당 → 선거·의회 → 정부 → 정책·권력 수정 법 당사자 → 소송·증거 → 법원 → 항소·재심 → 판결·법리 수정 **그리고 이 차이가 각 시스템의 성격을 만든다** 시장은: **행동 기반 신호** 가 강하다. 민주주의는: **집단선호의 압축** 이 강하다. 법은: **이유와 절차의 명시화** 가 강하다. 따라서 서로 대체하기 어렵다. **민주주의의 대표적 강점** 시민이: 이 정부를 더 이상 원하지 않는다. 는 신호를 보낼 수 있다. 하지만: 정확히 어느 정책 때문인가? 는 모호하다. 따라서: **POWER CORRECTION** 강함 **CAUSAL DIAGNOSIS** 거침 이다. **법의 대표적 강점** 법은: 왜 이 판단이 내려졌는가? 를 이유로 남긴다. 항소할 때도: 어떤 점이 잘못됐는가? 를 구체적으로 공격한다. 따라서: **ATTRIBUTION / REVIEW** 강함 이다. 반면 느리고 비용이 많이 든다. **세 제도를 같은 4층에 올려보면** • 층 — 시장 — 민주주의 — 법

• 탐색 — 기업·상품 경쟁 — 정책·정당 경쟁 — 주장·법리 경쟁 • 감지 — 가격·매출·손실 — 여론·투표·항의 — 소송·증거·항소 • 판별 — 기업분석·경쟁비교 — 공개논쟁·전문분석 — 이유제시·상급심 심사 • 수정 — 가격·상품·기업 퇴출 — 정책·정부 교체 — 판결·판례·법 개정 **바로 이것이 '인터페이스'의 핵심이다** 세상에 오류가 존재한다고 사회가 자동으로 그것을 알지 못한다. 오류는 반드시: **사회가 읽을 수 있는 신호** 로 변환되어야 한다. 시장에서는: **가격** 민주주의에서는: **표·여론** 법에서는: **주장·증거** 로 번역된다. 따라서: **REALITY → REPRESENTATION** 단계가 반드시 존재한다. **그리고 이 번역에서 정보가 손실된다** 시장가격 하나는:

- 환경피해

- 감정
- 정의

를 모두 담지 못한다. 선거 한 표는:

- 왜 그 후보를 골랐는지
- 어떤 정책을 지지했는지

를 모두 담지 못한다. 판결문 역시 현실 전체를 담지 못한다. 즉: **SIGNAL IS A COMPRESSION OF REALITY** 이다. 이것은 매우 중요한 **HARD LOCK**이다. 따라서 모든 오류수정 시스템에는 '표현손실'이 있다 Reality ↓ Sensor ↓ Signal 단계에서 이미 일부 정보가 사라진다. 그러므로: 센서가 무엇을 볼 수 있는가 가 제도의 성격을 결정한다. 시장: 가격화 가능한 것. 민주주의: 정치적으로 표현 가능한 것. 법: 법적으로 주장·입증 가능한 것. 이것이 세 제도의 사각지대가 생기는 이유다 환경오염이 가격에 반영되지 않으면 시장센서가 약하다. 말하지 못하는 소수자의 문제가 정치의제로 올라오지 않으면 민주주의센서가 약하다. 증거를 확보할 수 없는 피해는 법적 센서가 약하다. 따라서: 문제 존재 ≠ 제도적 가시성 이다. 따라서 학습루프의 실제 모습은 직선이 아니다 겉으로는: 탐색 → 감지 → 판별 → 수정 이지만 실제로는: 현실 ↓ 센서 ↓ 표현 ↓ 전달 ↓ 경쟁적 해석 ↓ 결정 ↓ 실행 ↓ 기억 ↓ 재측정 이다. 네 층은 이 복잡한 전달망을 기능적으로 압축한 것이다. 그래서 네 층의 정확한 경계를 최종 확정한다 LAYER 1 — 탐색 행동 이전 무엇을 시험할 것인가. LAYER 2 — 감지 결과 이후 / 해석 이전 무슨 차이가 생겼는가. LAYER 3 — 판별 신호 이후 / 결정 이전 그 차이를 무엇 때문이라고 설명할 것인가. LAYER 4 — 수정 설명 이후 / 다음 행동 이전 어떤 모델·행동·규칙을 실제로 바꿀 것인가. 이 시간적 경계가 중요하다. 행동 → 결과 → 해석 → 재행동 의 네 구간과 정확히 대응한다. 최종 CANONICAL FORM 문명 학습의 네 층 1. 탐색 가능한 대안 → 현실 시험 2. 감지 현실 결과 → 오류신호 3. 판별 오류신호 → 수정할 원인/모델 4. 수정 판단 → 새로운 행동·규칙·기억 그리고 층과 층 사이에는 반드시:

전달 인터페이스 가 존재한다. 따라서 현재 가장 강한 전체식 VARIATION ↓ EXPOSURE ↓ OUTCOME ↓ SENSING ↓ SIGNAL ↓ TRANSMISSION ↓ ATTRIBUTION ↓ CONTEST ↓ AUTHORITY ↓ UPDATE ↓ RETENTION ↓ RE-EXPOSURE 이다. 이를 인간이 이해하기 쉽게 압축하면: 탐색 → 감지 → 판별 → 수정 이다. 최종 시장·민주주의·법 비대칭 시장 “사람들이 실제로 선택했는가?” 강점: 빠른 행동 피드백 주된 수정: EXIT / PRICE / REALLOCATION 민주주의 “시민들이 이 통치를 계속 위임하는가?” 강점: 권력의 교체 가능성 주된 수정: VOICE / VOTE / REPLACEMENT 법 “이 판단은 이유·증거·권리·절차를 견디는가?” 강점: 이유에 기반한 재검토 주된 수정: CHALLENGE / APPEAL / REVISION 따라서: 시장은 선택을 시험하고, 민주주의는 권력을 시험하며, 법은 판단을 시험한다. 이 세 문장이 현재 비대칭을 가장 짧게 표현한다. MASTER HARD LOCKS v1.1 대안 존재 ≠ 탐색 결과 발생 ≠ 감지 감지 ≠ 원인판별 판별 ≠ 수정 현실의 오류 ≠ 제도적 오류신호 신호 ≠ 현실 전체 오류신호 도달 ≠ 수정권한 존재 정책 변경 ≠ 학습 시장가격 ≠ 사회적 가치 투표 ≠ 정확한 정책평가 판결 ≠ 최종 진실 최종 압축문 문명은 단순히 실패를 경험함으로써 학습하지 않는다. 현실에서 발생한 차이가 센서에 포착되고, 제도가 읽을 수 있는 신호로 번역되고, 결정권자에게 전달되고, 경쟁하는 설명 속에서 원인이 판별되고, 실제 규칙과 행동이 수정되어 다시 현실에 노출될 때 비로소 학습루프가 닫힌다.

그리고 시장·민주주의·법은 이 하나의 루프를 복제한 세 장치가 아니다. 시장은 선택의 실패를,

민주주의는 권력위임의 실패를, 법은 판단·권리·절차의 실패를 서로 다른 센서와 전달경로를 통해 탐지한다. 따라서 강한 문명은 하나의 완벽한 평가기관을 가진 문명이라기보다, 서로 다른 현실을 감지하는 여러 오류수정망이 겹쳐 있고, 한 시스템의 맹점을 다른 시스템이 다시 감지할 수 있는 문명이라고 정의하는 편이 더 정확하다. 이 버전부터는 탐색→감지→판별→수정을 단순 네 단어가 아니라 **행동 이전 / 결과 이후 / 해석 / 제도변화**라는 시간적 경계까지 포함한 정본 구조로 써도 되겠습니다. 되먹임은 왜 세상을 지능처럼 보이게 만드는가 제어이론 · 항상성 · 정보 · 예측오차 · 강화학습 · 진화 무엇이 진짜 같은 수학이고 무엇이 위험한 비유인가 여러 영역에서 실제로 공유되는 수학적 골격은 있다. 특히: **상태 → 관측 → 차이 → 행동 → 새로운 상태** 라는 **페루프 구조**는 제어이론·강화학습·생리적 조절 등에서 실제 형식적으로 연결된다. 적응제어와 강화학습 역시 불확실한 동적 시스템을 제어한다는 공통 문제를 서로 다른 방식으로 푼다. 그러나: **피드백 ≠ 학습 ≠ 진화 ≠ 지능 ≠ 목적** 이다. 공통점보다 **상태가 어디에 저장되고, 무엇이 비교되며, 어떤 메커니즘이 다음 반응을 바꾸는가**를 구분해야 한다. **먼저 왜 모든 것이 닮아 보이는가** 온도조절기. 혈당조절. 바둑 AI. 동물 학습. 자연선택. 시장. 민주주의. 겉으로 보면 모두 이런 모습을 갖는다. 현재 상태 ↓ 무언가 감지 ↓ 차이가 생김 ↓ 반응 ↓ 새로운 상태 그래서 매우 강한 유혹이 생긴다. 세상은 모두 하나의 거대한 피드백 알고리즘 아닐까? 사이버네틱스 자체도 역사적으로 동물·기계·조직의 제어와 의사소통을 공통된 피드백 개념으로 연구하려는 시도였다. 하지만 이제부터는: **닮아 보이는 구조와 동일한 작동원리**를 분리해야 한다. 여기에는 '학습'이 없어도 된다 중요하다.

온도조절기가 매일 똑같은 규칙을 쓴다고 하자. 온도가 낮으면 난방 ↑ 입니다. 어제와 같은 규칙. 오늘도 같은 규칙. 내일도 같은 규칙. 그러면 시스템은: **피드백 제어**는 하지만 **학습**하지 않을 수 있다. 왜냐하면 과거 경험 때문에: **제어규칙 자체가 바뀌지 않았기 때문**이다. **HARD LOCK 07 FEEDBACK ≠ LEARNING** 피드백은 현재 오차를 이용해 행동을 바꿀 수 있다. 학습은 과거 경험 때문에: **미래의 반응규칙이 달라져야 한다. 적응제어는 그 중간에 있다** 온도조절 시스템이 오래 쓰다 보니 건물의 열손실 특성을 학습한다고 해보자. 예전에는: 난방 10분이면 2도 오른다. 고 생각했다. 그런데 실제로는 겨울이 되면서 달라진다. 시스템이 모델의 매개변수를 수정한다. 그러면: **controller 자체가 환경에 맞게 변한다.** 이것이 **적응제어**의 영역이다. 적응제어와 강화학습은 모두 불확실한 시스템의 제어문제를 다루지만, 전자는 특정 구조를 전제하고 실시간 안정성·매개변수 적응에 강한 반면, 강화학습은 더 넓은 문제에서 경험을 통해 정책을 학습하는 데 강점이 있다는 비교가 가능하다. 그래서 **지능의 핵심도 '미래를 완벽히 맞히는 것'이 아닐 수 있다** 기존 그림: 세상을 완벽히 이해한다. ↓ 완벽한 행동을 한다. 더 현실적인 그림: 대략적인 모델로 행동한다. ↓ 결과를 본다. ↓ 계속 수정한다. 즉: **PREDICTION QUALITY + CORRECTION SPEED** 둘 다 중요하다. 이것은 **목적함수와 같으면서도 다르다** 혈당이 일정 범위에서 유지된다고 해서: 몸이 혈당 90을 철학적으로 원한다. 고 말할 필요는 없다. 조절회로가:

- 센서
- 호르몬
- 대사기관

을 통해 범위를 유지하기 때문이다. 즉: **REGULATION ≠ CONSCIOUS GOAL** 이다. **HARD LOCK 08 SET POINT ≠ DESIRE** 생물학적 조절값을 인간의 의도와 혼동하면 안 된다. 이제 뇌로 가자 **PREDICTION ERROR** 뇌를 설명하는 여러 이론 중에는: 뇌가 환경의 상태를 예측하고 감각입력과 예측의 차이를 이용해 내부모델을 수정한다. 는 계열이 있습니다. 아주 단순화하면: prediction vs

observation ↓ **prediction error** ↓ model update 입니다. 여기서는 문명 루프의: 감지 → 판별 → 수정 과 상당히 닮아 보입니다. **강화학습은 훨씬 명확한 비교대상이다** 강화학습에서는 행위자가: 상태 s 에서 행동 a 를 합니다. 보상: r 을 얻습니다. 그리고: 이 행동이 장기적으로 얼마나 좋은가? 를 학습합니다. 즉: state → action → reward → update → new policy 입니다. 강화학습은 실제로 환경과의 상호작용에서 평가적 피드백을 이용해 행동결정을 개선하는 학습문제로 정의됩니다. **그리고 바둑 AI와 바로 연결된다** 바둑 AI는:

- 자기대국
- 승패
- 가치평가
- 정책 갱신

등을 통해 더 강해집니다. 즉 한 판의 결과가: **다음 판의 행동확률** 을 바꿉니다. 여기에는 명확한: **MEMORY OF EXPERIENCE** 가 있습니다. 그러나 **진화와 학습은 동일하지 않다** 이 차이를 강하게 잡아야 한다. 개체학습: 한 개체 가 경험하고 ↓ 그 개체 내부가 변하며 ↓ 그 개체의 이후 행동이 바뀐다. 진화: 여러 개체의 변이 가 있고 ↓ 번식 성공이 다르며 ↓ **집단의 구성비** 가 바뀐다. 즉 변화가 저장되는 위치가 다르다. 따라서 '**진화가 학습한다**'는 문장은 두 수준으로 나뉜다 약한 의미 환경에 따라 다음 세대의 구조가 체계적으로 달라진다. **ACCEPT** 강한 의미 진화가 환경을 모델링하고 오류를 계산해서 자신의 정책을 의식적으로 수정한다. **REJECT** 이다. **HARD LOCK 09 POPULATION UPDATE** ≠ **AGENT LEARNING** 구조적으로 대응할 수 있지만 메커니즘은 다르다. 그런데 이것만으로는 너무 넓다 돌이 산에서 굴러잡니다. 현재 위치: xt 힘이 작용합니다. 다음 위치: xt+1 입니다. 그러면 방금 구조에 들어잡니다. 그렇다고 돌이 학습했다고 할 수는 없습니다. 따라서 조건 하나가 더 필요합니다. 이것이 작동조건으로서의 학습이다 따라서: **LEARNING OPERATIONAL DEFINITION v1.0** 환경과의 과거 상호작용이 시스템 내부 또는 집단적으로 보존되는 상태변화를 만들고, 그 상태변화가 이후 유사한 조건에서 반응확률·정책·구조를 체계적으로 변화시키면 그 시스템은 **최소 의미에서 학습한다**. 핵심은 세 가지다. **RETENTION** 흔적이 남는다. **TRANSFER** 다음 상황에 영향을 준다. **SYSTEMATIC CHANGE** 단순 노화·손상이 아니라 경험과 연결된 방식으로 반응이 달라진다. 이제 '오류'를 다시 정의할 수 있다 앞 장에서는: 기대 또는 기준과 결과 사이에서 수정 가능한 정보가 된 차이 라고 정의했습니다. 이를 더 일반화하면:

ERROR SIGNAL OPERATIONAL DEFINITION v1.0 시스템이 유지·달성·예측하려는 기준과 관측된 결과 사이의 차이 가운데, 그 차이를 이용했을 때 이후 조절·모델·정책을 바꿀 수 있는 정보가 되는 것. 즉: **Difference + Reference + Actionability** 세 가지가 필요합니다. 이것이 중요한 이유 진화에서: 잘못된 개체 라는 것이 존재하는가? 엄밀히는 아닙니다. 어떤 환경에서 번식 성공이 낮은 변이가 있을 뿐입니다. 그러므로: **evolutionary failure** 와 **prediction error** 를 완전히 같은 의미의 ERROR로 부르면 안 됩니다. 그리고 인간·사회에는 **TYPE D**가 있다 **TYPE D — NORMATIVE ERROR** 어떤 가치·규칙·권리 기준 vs 실제 상태 예: 누구나 적법절차를 받아야 한다. 그런데 실제로 절차가 지켜지지 않았다. 여기서는 자연적 사실만으로 오류가 정해지지 않는다. **규범** 이 필요하다. 그러면 정보는 어디에 들어가는가 정보라는 말을: 지식 과 동일시하면 안 됩니다. 정보이론에서는 훨씬 더 형식적인 개념입니다. 핵심 직관은: **불확실성을 얼마나 줄이는가** 입니다. 따라서 좋은 센서는: 데이터가 많다. 가 아니라 다음 상태를 구별하는 데 유용한 차이를 만든다 가 중요합니다. 그러므로 정보와 학습의

관계도 분리한다 INFORMATION ≠ LEARNING 정보가 들어왔다. 그러나 정책이 안 바뀐다. 학습이 아닙니다. 정보가 들어와: modelt → modelt+1 이 되고, 미래 행동이 달라져야: 학습 입니다. **여기서 매우 중요한 반례가 나온다** 진화에는 인간 문명의 의미에서: **판별** 이 없을 수 있습니다. 자연선택은: 이 개체가 죽은 이유는 온도 때문이었다. 라고 분석하지 않습니다. 차등번식이 발생하고 집단구성이 변할 뿐입니다. 즉 문명의: 탐색 → 감지 → **판별** → 수정 을 그대로 자연선택에 투영하면 틀립니다.

이 발견 때문에 MASTER를 두 계열로 갈라야 한다 FAMILY A — REGULATORY LOOP 상태 → 측정 → 차이 → 반응 대표:

- 온도조절
- 항상성
- 단순 제어

FAMILY B — LEARNING LOOP 행동/예측 → 결과 → 정보 보존 → 모델/정책 수정 대표:

- 인간
- 강화학습
- 과학
- 문명

FAMILY C — SELECTION LOOP 변이 → 환경 노출 → 차등 존속/번식 → 집단 구성 변화 대표:

- 자연선택
- 일부 시장 경쟁

이게 상당히 중요한 그림이다 예를 들어 인간: 초 단위 혈압·체온·자세 조절 REGULATION 하루~평생 경험·습관·학습 LEARNING 세대 유전적 변화 EVOLUTION 입니다. 즉: SAME ORGANISM DIFFERENT UPDATE TIMESCALES 입니다. **문명에서도 세 시간척도가 보인다** 즉각 가격·금리·행정 대응. REGULATION 수년 정책·기업·정치학습. INSTITUTIONAL LEARNING 세대~세기 제도·문화·기술생태계의 선택과 축적. CULTURAL EVOLUTION 이다. **이것은 바둑에도 있다** 한 수: 국소 전술. 20수: 전투. 한 판: 전략. 100판: 기풍과 습관. 즉: MOVE ERROR STRATEGY ERROR POLICY ERROR 가 다릅니다. **사회에서도 똑같다** 기업이 한 분기 손실을 봤습니다. 하지만 장기 연구개발 때문이라면? 시장센서는: 단기 손실 을 보여줍니다. 그러나 전략적으로는: 장기 투자 일 수 있습니다. 정부가 단기 지지율을 잃었지만: 장기적으로 필요한 개혁일 수 있습니다. 따라서 오류판단에는: HORIZON 이 반드시 들어가야 합니다. **이것은 문명학습 이론을 훨씬 강하게 만든다** 같은 정책도: 개인 수준에서는 손해. 산업 수준에서는 이익. 현재세대에는 이익. 미래세대에는 손해. 일 수 있다. 그러므로: **누구의 오류인가?** 를 묻지 않고: 사회가 잘못했다. 고 말하면 너무 거칠다. **이것이 시장·민주주의·법의 비대칭이 중요한 진짜 이유다** 서로 다른 제도가: 같은 것을 세 번 측정 하는 것이 아니다. 각자가: 다른 오류공간 을 측정합니다. 시장: 경제적 실행가능성. 민주주의: 정치적 정당성·위임. 법: 권리·절차·규칙 정합성. 과학: 경험적 사실과 설명. 따라서 문명은: **다중센서 시스템** 으로 보아야 한다. **그래서 문명의 최종 알고리즘은 단순 학습루프가 아니다**

더 정확히는: 여러 센서 ↓ 여러 오류신호 ↓ 서로 다른 모델 ↓ 가치충돌 ↓ 집단적 선택 ↓ 결과 ↓ 재평가 입니다. 즉: MULTI-LOOP GOVERNANCE 에 가깝습니다. **그렇다면 왜 세상은 지능처럼**

보이는가 이 질문에 이제 꽤 좋은 답을 줄 수 있습니다. 피드백이 있기 때문만은 아니다. 더 정확히는 여러 단계의 시스템에서:

- 상태가 결과를 낳고
- 결과가 다음 상태에 영향을 주며
- 일부 시스템에서는 그 영향이 기억되고
- 일부에서는 정책이 바뀌며
- 일부에서는 구조가 선택되고

이 과정이 반복되기 때문입니다. 즉: **ITERATED STATE-DEPENDENT UPDATE** 가 광범위하게 존재합니다. **MASTER MODEL v0.8 THREE LOOPS LOOP 1 — REGULATION** 틀어지면 되돌린다. 상태 → 차이 → 교정 대표: 온도조절·항상성. **LOOP 2 — LEARNING** 틀린 경험 때문에 다음번에는 다르게 한다. 예측/행동 → 결과 → 기억 → 모델/정책 수정 대표: 뇌·인간·강화학습·과학. **LOOP 3 — SELECTION** 서로 다른 구조가 다른 성공률을 보여 집단구성이 달라진다. 변이 → 환경 → 차등 존속 → 구성변화 대표: 진화·일부 시장경쟁. 이 장의 가장 중요한 수정 이전의 큰 명제: 생명·뇌·AI·문명은 모두 ERROR → UPDATE 구조다. 는 너무 넓었다. 정본은 이렇게 수정한다. 생명·AI·문명에는 결과가 다음 상태를 바꾸는 반복적 폐루프가 널리 존재한다. 그러나 그 폐루프는 최소한 조절·학습·선택이라는 서로 다른 메커니즘으로 나뉘며, '오류'의 기준·기억의 위치·수정 메커니즘이 다르다. 따라서 구조적 상동성을 인정하되 하나의 보편적 학습 알고리즘으로 환원하지 않는다. **HARD LOCKS v0.8 피드백 ≠ 학습 조절 ≠ 학습 학습 ≠ 진화 진화 ≠ 의도적 최적화 정보 ≠ 지식 정보 입력 ≠ 학습 차이 ≠ 오류 적합도 차이 ≠ 예측오차 변화 ≠ 적응 적응 ≠ 지능 수학적 최적화 표현 ≠ 자연의 목적 국소 오류 ≠ 전체 시스템 오류** 그리고 이제 아주 중요한 다음 질문이 열린다 세 종류의 루프를 구분했습니다. 그런데 인간에게는 이상한 능력이 하나 더 있습니다. 온도조절기는: 왜 22도를 유지해야 하지? 라고 묻지 않습니다. 진화는: 왜 번식해야 하지? 라고 묻지 않습니다. 바둑 AI도: 왜 승리해야 하지? 라고 묻지 않습니다. 그러나 인간은: 자신에게 주어진 목적함수 자체를 대상으로 만들 수 있다. 왜 돈을 벌려고 하는가?

왜 이겨야 하는가? 왜 살아야 하는가? 무엇을 좋은 삶이라고 부를 것인가? 그러므로 인간에게는 세 루프 위에 하나를 더 올려야 할 가능성이 생겼습니다. **LOOP 4 — REFLECTION** 무엇을 수정해야 하는지가 아니라 무엇을 목표로 삼아야 하는지를 수정한다. 가치 → 행동 → 경험 → 가치에 대한 반성 → 가치 수정입니다. 이것이 실제로 독립적인 네 번째 루프인지, 아니면 고차원 학습의 한 종류일 뿐인지는 아직 검증하지 않았습니다.

5장. 반성과 자기모델

인간은 왜 자신이 무엇을 원하는지 다시 물을 수 있는가?

반성적 행위성은 현재의 선호를 실행하는 데서 그치지 않고 선호·목표·평가기준을 검토하고 수정하는 능력이다. 자기모델은 이 검토를 도울 수 있지만 의식의 증거와 같지는 않다. 좋습니다. 이번에는 REFLECTION을 멋진 개념어로 남기지 않고, 사례를 통과해야만 인정되는 작동조건으로 바꾸겠습니다. 동시에 네 루프를 “상위/하위”로 잘못 쌓지 않고, 서로 다른 시간척도·변경대상·기준의 층으로 재정리하겠습니다. 사례 기반 판별 · 네 루프 위계 재정리 · 메타인지와 반성의 엄격한 분리 TARGET: REFLECTION의 독립 작동조건과 네 루프의 관계 먼저 가장 중요한 수정 기준에는: REGULATION LEARNING SELECTION REFLECTION 을 나란히 놓았다. 하지만 이 네 가지는 같은 종류의 층이 아니다. 특히: SELECTION 은 REGULATION이나 LEARNING보다 “한 단계 위”가 아니다. 전혀 다른 업데이트 메커니즘이다. 그리고: REFLECTION 은 그 셋 위에 동일한 방식으로 놓이는 네 번째 자연적 루프도 아니다. 따라서 위계를 다시 짜야 한다. 네 루프를 먼저 ‘무엇이 바뀌는가’로 구분한다 LOOP A — REGULATION 현재 상태를 바꾼다 질문: 지금 상태를 목표범위로 어떻게 되돌릴까? 변경대상: 현재 출력 / 현재 상태 예:

- 체온
- 혈압
- 차량 방향
- 난방기 출력

구조: 차이 감지 → 즉각 조절 → 상태 복귀 LOOP B — LEARNING 미래 반응규칙을 바꾼다 질문: 지난 경험 때문에 다음번에는 무엇을 다르게 할까? 변경대상: 모델 / 정책 / 반응확률 예:

- 바둑기사의 기풍 수정
- 강화학습 정책
- 사람의 습관 변화
- 기업 전략 수정

구조: 경험 → 기억 → 모델/정책 수정 → 다음 행동 변화 LOOP C — SELECTION 집단의 구성비를 바꾼다 질문: 서로 다른 변이 가운데 무엇이 더 많이 남는가? 변경대상: 집단 구성 예:

- 자연선택
- 기업의 진입·퇴출
- 문화적 관행의 확산과 소멸

구조: 변이 → 차등 성공 → 보존·확산 차이 → 구성 변화 여기에는 반드시:

- 자기모델

- 원인판별
- 의식적 수정

이 필요하지 않다. *LOOP D — REFLECTION* 무엇을 수정할지 정하는 기준 자체를 바꾼다 질문: 내가 왜 이것을 목표로 삼고 있는가? 변경대상: **가치 우선순위 / 상위 목표 / 자기정책의 정당화 기준** 예:

- 돈 극대화에서 시간·건강 우선으로 삶의 목표 변경
- 승리 중심의 바둑관에서 창작·교육 중심으로 방향 전환
- 조직이 매출보다 안전을 상위 기준으로 재설정

구조: 현재 가치/목표 → 자기·규범 평가 → 기준 충돌 발견 → 가치 우선순위 수정 → 새 정책 **따라서 네 루프의 관계는 피라미드가 아니다** 잘못된 그림: REGULATION ↓ LEARNING ↓ SELECTION ↓ REFLECTION 이 아니다. 더 정확한 구조는: **REGULATION** 현재 상태 조절 **LEARNING** 개체·시스템의 정책 수정 **SELECTION** 집단 구성 변화 이 세 가지가 서로 다른 시간척도와 저장위치에서 작동하고, **REFLECTION** 은 일부 학습 가능한 행위자에게서: **학습의 기준 자체를 대상으로 삼는 메타층** 으로 붙는다. 즉: REGULATION 과 LEARNING 은 한 시스템 안에서 연속될 수 있고, SELECTION 은 그 시스템들 사이의 집단 수준에서 작동할 수 있으며, REFLECTION 은 일부 LEARNING 시스템에서만 나타날 수 있다. 네 루프를 '무엇을 수정하는가'로 한 번에 비교하면 • 루프 — 수정되는 것 — 기억 필요 — 내부 기준 필요 — 자기모델 필요 • Regulation — 현재 상태 — 필수 아님 — 있음 — 아님 • Learning — 미래 정책·모델 — 필요 — 있음 — 필수 아님 • Selection — 집단 구성 — 개체기억 불필요 — 내부 기준 불필요 가능 — 아님 • Reflection — 가치·목표·자기정책 기준 — 필요 — 필요 — 강하게 필요 이 표가 현재 가장 정확한 위계다. **CASE 4** 다이어트 사람이 말한다. 야식을 먹으면 다음날 힘드니까 끊어야겠다. 이것은: **LEARNING** 이다. 현재 가치: 건강해지고 싶다. 는 그대로다. 행동정책만 수정한다. **CASE 7** “우리가 매출을 키우려고 고객 중독을 만들고 있는 것 아닌가?” 회사가:

- 체류시간
- 클릭률
- 매출

을 계속 최적화했다. 그런데 내부 논쟁 끝에: 이용자의 장기적 복지를 손상시키는 성장은 좋은 성장이 아니다. 라고 결정한다. 이후 KPI를: 체류시간 에서 장기적 만족 + 안전 + 신뢰 로 일부 변경한다. 이것은: **REFLECTION** 이다. 왜냐하면: **보상함수 일부를 다시 정의했기 때문이다.** **REFLECTION**의 첫 번째 판별기준 질문: 목표는 그대로인데 방법만 바뀌었는가? YES: **LEARNING** 이다. NO: 목표·가치기준 자체가 바뀌었다면 다음 검사를 한다. 두 번째 판별기준 질문: **단순히 욕망이 바뀐 것인가, 아니면 기존 욕망을 평가해 수정한 것인가?** 예: 갑자기 돈에 흥미를 잃었다. 이것은 단순: **preference change** 일 수 있다. 반면: 돈을 좇아보니 내가 원하는 삶과 충돌한다는 것을 깨달았고, 그래서 돈의 우선순위를 낮추겠다. 라면: **reflection** 후보다. 즉: **VALUE CHANGE ≠ REFLECTION** 이다. 반드시: **평가를 통한 가치변경** 이 있어야 한다. **REFLECTION OPERATIONAL TEST v1.0** REFLECTION을 인정하려면 다섯 조건이 필요하다. *R1 — 자기대상화* 자신의:

- 욕망
- 목표

- 가치
- 정책

중 하나를 대상으로 삼는다. R2 — 상위기준 비교 현재 상태를:

- 장기적 가치
- 규범
- 정체성
- 다른 가치

와 비교한다. R3 — 이유 평가 단순 감정변화가 아니라: 왜 유지하거나 바꿀 것인가? 에 대한 이유구조가 존재한다. R4 — 기준 수정 행동수단뿐 아니라: 가치 우선순위 또는 상위 정책기준 이 실제로 바뀐다. R5 — 미래전이 그 변경이 이후 유사한 선택에 반복적으로 영향을 준다. 다섯 조건을 모두 통과해야: REFLECTION LOOP COMPLETE 로 인정한다. 이제 메타인지와 반성을 분리한다 여기서 가장 중요한 보정이다. 기존에는 메타인지와 reflection이 너무 가까이 붙어 있었다. 둘은 다르다. 반성의 최소 정의 REFLECTION 은 이번 프로젝트에서: 자신의 욕망·목표·가치·자기정책을 이유와 상위기준에 비추어 평가하고, 필요하면 그 기준 자체를 수정하는 것 이다. 대상은: valuation / goal / self-policy 이다. 그래서 둘의 질문이 다르다 메타인지: “내가 제대로 알고 있는가?” Reflection: “내가 이것을 중요하게 여기는 것이 타당한가?” 이다. 의학 사례 의사: 이 진단에 대한 내 확신이 너무 높은 것 같다. 감별진단을 다시 보자. METACOGNITION 의사: 검사수를 늘려 진단률을 높이고 있지만 환자에게 불필요한 부담을 만들고 있다. 내가 무엇을 '좋은 진료'라고 정의하고 있는지 다시 봐야 한다. REFLECTION 이다. 반대로 REFLECTION에는 메타인지가 어느 정도 필요하다 자신의 가치와 욕망을 평가하려면 최소한: 내가 지금 무엇을 원하고 있는가? 를 어느 정도 알아차려야 한다. 따라서: METACOGNITION / SELF-MONITORING 은 reflection의 일부 기반이 될 수 있다. 하지만: 필요조건 후보 ≠ 충분조건 이다. 메타인지도 두 종류로 나누면 더 선명하다 EPISTEMIC METACOGNITION 내가 무엇을 얼마나 아는가? 예: 확신도, 기억 신뢰도. CONTROL METACOGNITION

지금 더 생각해야 하는가, 멈춰야 하는가? 예: 추가 검사 여부, 탐색시간 배분. 이 둘은 모두: 판단과 정보처리의 질 을 다룬다. 아직: 가치 자체 를 다루지 않는다. 네 루프를 시간척도까지 넣어 다시 배열한다 MICRO — REGULATION 초~분. 현재 상태를 조절. MESO — LEARNING 반복경험~생애. 정책·모델 수정. POPULATION / GENERATIONAL — SELECTION 세대·집단. 구성 변화. META — REFLECTION 특정 고차 학습시스템 내부에서 필요 시 발생. 시간척도는 고정되지 않는다. 몇 초 만에 일어날 수 있고 수십 년 걸릴 수도 있다. 따라서 REFLECTION은: 시간적 상위 가 아니라 논리적 상위 다. 수정대상이 상위기준이기 때문이다. 이것이 최종 위계다 물리적/생리적 층 REGULATION 현재 상태를 바꾼다. ↓ 행위자 층 LEARNING 미래 정책을 바꾼다. 그리고 별도의 집단층: POPULATION AXIS SELECTION 어떤 정책·구조를 가진 시스템이 더 많이 남는지 바꾼다. LEARNING 시스템 가운데 일부에는: META-LAYER REFLECTION 정책을 평가하는 가치기준을 바꾼다. 즉: REGULATION → LEARNING 에는 어느 정도 기능적 계층성이 있지만, SELECTION 은 옆의 집단층이고, REFLECTION 은 LEARNING 위의 메타층이다. 이것이 가장 정확하다. 여기서 아주 중요한 사례 — 중독 약물 사용. 단계 1 약물 보상에 따라 행동 반복. LEARNING 단계 2 사람이 생각한다. 약물을 원하고 있다. 이것은: SELF-MONITORING / METACOGNITION 일 수 있다. 단계 3 하지만 나는 이 욕망이 내 삶을 지배하는 것을 원하지 않는다. 여기서: VALUE CONFLICT 가 생긴다. 단계 4 건강과

가족관계를 더 상위 가치로 두겠다. 가치 우선순위를 평가한다.

REFLECTION 단계 5 치료를 받고 환경을 바꾸고 접근을 차단한다. **LEARNING / SELF-POLICY IMPLEMENTATION** 이다. 즉 실제 인간에서는: **REFLECTION**과 **LEARNING**이 연속적으로 연결된다. Reflection이 기준을 바꾸고, Learning이 그 기준을 행동정책으로 구현한다. 이것이 중요한 발견이다 Reflection만으로 행동이 바뀌지 않는다. 알지만 못 한다. 가 가능합니다. 따라서: **REFLECTION** ≠ **SELF-CONTROL** 이다. Reflection은: 무엇을 바꿀지 결정하는 층 이고, Learning / regulation은: 실제 행동을 바꾸는 구현층 이다. 따라서 **REFLECTION**에도 전달경로가 있다 현재 목표 ↓ 목표 충돌 경험 ↓ 자기/사회적 가시화 ↓ 상위기준 제시 ↓ 이유 경쟁 ↓ 가치 우선순위 변경 ↓ 정책 번역 ↓ 행동 변화 ↓ 재평가 이다. 즉 Reflection도: **생각 한 번** 이 아니라 페루프다. **REFLECTION 실패도 나눌 수 있다 FAILURE 1** — 자기비가시성 내가 무엇을 원하는지 모른다. **FAILURE 2** — 비교기준 부재 욕망을 평가할 다른 가치가 없다. **FAILURE 3** — 정당화 실패 단순 감정변화를 이유로 착각한다. **FAILURE 4** — 가치 수정 실패 문제를 알아도 기존 가치위계를 못 바꾼다. **FAILURE 5** — 실행 번역 실패 가치는 바뀌었지만 행동정책이 그대로다. 반대로 생각이 많아도 반성이 아닐 수 있다 밤새 고민한다. 어떻게 하면 더 많이 벌까? 어떤 방법이 더 효율적일까? 아무리 복잡하게 생각해도: 돈 최대화 라는 목표를 전혀 검토하지 않았다면 그것은: **sophisticated optimization** 이지 **reflection** 은 아니다. 그렇다고 AI가 **REFLECTION**을 영원히 못한다고 말하면 안 된다 어떤 AI가:

- 자신의 목표를 표현하고
- 서로 충돌하는 가치를 비교하고
- 이유를 평가하고
- 상위 가치기준을 수정하며
- 수정된 기준을 이후 행동에 지속적으로 반영

한다면? 현재 이 프로젝트의 작동정의에 따르면: 기능적으로 **REFLECTION** 유사 루프 를 인정해야 할 가능성이 있다. 따라서: **REFLECTION: 인간 본질** 로 잡그지 않는다. 더 정확한 인간의 특수성 후보 현재 인간에게 특히 강하게 보이는 것은: 언어화된 자기모델 + 사회적으로 공유되는 규범 + 장기 자기서사 + 가치의 공개적 정당화 를 이용한 reflection이다. 즉: 나는 왜 이렇게 살아야 하지? 라는 질문을 혼자뿐 아니라:

- 철학
- 종교
- 법
- 정치
- 문학

을 통해 집단적으로 수행합니다. 네 루프 최종 구조 v1.0 I. **REGULATION** — 상태수정 “지금 바로 무엇을 고칠까?” 현재 상태 → 오차 → 교정 II. **LEARNING** — 정책수정 “다음번에는 무엇을 다르게 할까?” 경험 → 기억 → 모델/정책 갱신 III. **SELECTION** — 구성수정 “어떤 변이·정책·구조가 더 많이 남는가?” 변이 → 차등 성공 → 구성 변화 ※ I-II보다 상위 단계가 아니라 다른 집단수준 메커니즘. IV. **REFLECTION** — 기준수정 “왜 이것을 좋은 결과라고 부르는가?” 가치/목표 → 자기·규범 평가 → 가치 우선순위 수정 → 새 정책 ※ II 위의 메타수준으로만 배치. 네 루프를 가장 짧게 기억하면 **REGULA-**

TION 출력을 고친다. LEARNING 정책을 고친다. SELECTION 구성을 바꾼다. REFLECTION 평가기준을 고친다. 이 네 문장이 현재 가장 선명하다. 최종 판별표 · 사례 — 메타인지 — 학습 — Reflection · “이 답에 확신이 낮다” — O — X — X · “이 수읽기에 확신이 낮으니 더 읽자” — O — 일부 — X · “이 전략은 자꾸 실패하니 바꾸자” — 가능 — O — X · “나는 왜 승리만 가치로 두고 있지?” — O — 아직 X — O 시작 · “승리보다 창조성도 중요하게 두겠다” — O — 후속 필요 — O · “그 가치변경에 맞춰 실제 기풍·훈련을 바꾼다” — O — O — O 완료 이번 패치의 최종 판정 REFLECTION은 살아남는다. 하지만 정의를 좁힌다. REFLECTION은 자신의 인지상태를 관찰하는 메타인지가 아니다. 자신의 욕망·목표·가치·자기정책을 상위 이유와 규범에 비추어 평가하고, 그 결과 평가기준 자체를 수정하여 이후 행동정책에 지속적으로 반영하는 메타수준의 학습과정이다. 따라서: 생각함 ≠ 반성 자기관찰 ≠ 반성 선호변화 ≠ 반성 전략변경 ≠ 반성 가치기준의 평가·수정·전이: REFLECTION

으로 잠근다. 그리고 여기서 다음 질문이 더 정확해졌다 이제 묻고 싶은 것은: 인간은 왜 반성할 수 있는가? 보다 더 정밀하다. 어떤 시스템이 '자신의 판단기준'을 다시 하나의 대상으로 만들 수 있는가? 이 질문은:

- 자기모델
- 언어
- 재귀
- 장기기억
- 사회적 규범
- counterfactual reasoning

을 요구할 가능성이 있다. 따라서 다음 장는: 모델이 자기 자신을 모델 안에 넣는 순간 무엇이 달라지는가 세계모델 → 자기모델 → 메타인지 → 규범적 reflection 으로 가는 것이 가장 자연스럽다. 이제야 우주는 언제부터 자기 자신을 모델링하기 시작했는가라는 큰 질문을 과장 없이 다룰 준비가 됐다. 좋습니다. 이제 “자기모델”을 너무 쉽게 의식이나 자아와 동일시하지 않고, 세계모델 → 자기모델 → 메타인지 → 규범적 반성이 실제로 어떤 경계를 넘을 때 생기는지 판별하겠습니다. 모델이 자기 자신을 모델 안에 넣는 순간 무엇이 달라지는가 세계모델 · 자기모델 · 메타인지 · 재귀 · 규범적 반성 먼저 바둑기사에게 가장 단순한 질문 기사가 판을 봅니다. 그가 알아야 하는 것은 무엇입니까?

- 어디에 흑돌이 있는가
- 어디에 백돌이 있는가
- 어느 돌이 약한가
- 어디가 큰가

즉: 세계 상태 입니다. 이것만 잘 알아도 바둑을 둘 수 있을까요? 아닙니다. 기사에게는 또 하나가 필요합니다. 나는 이 변화를 얼마나 정확히 읽을 수 있는가? 남은 시간이 몇 분인가? 나는 복잡한 싸움에서 강한가? 지금 피곤한가? 즉 판뿐 아니라: 자기 자신도 상태변수가 됩니다. 여기서 자기모델의 문이 열립니다. 세계모델의 최소 정의 WORLD MODEL 은: 환경의 현재 상태와 변화규칙을 내부적으로 표현하여, 가능한 행동 이후의 결과를 예측하는 데 사용하는 구조 라고 둘 수 있습니다. 예: 바둑: 현재 판 → 이 수를 두면 → 상대가 이렇게 대응할 가능성 입니다. 자동차 운전: 앞차 속도 + 거리 + 도로상태 → 브레이크를 밟으면 어떻게 될까 입니다. 핵심: WORLD MODEL은 “나”를

반드시 포함할 필요가 없다. 자기모델은 철학적 '자아'보다 훨씬 작게 시작할 수 있다 이것이 중요합니다. 자기모델이라고 해서 곧바로: 나는 누구인가? 를 묻는 것은 아닙니다. 매우 작은 자기모델도 가능합니다. 예: 로봇팔이:

- 자신의 관절 위치
- 최대 회전각
- 현재 배터리

를 추정합니다. 기능적으로는: 자기 상태에 관한 모델 이 있습니다. 하지만 이것만으로: 자아의식 이 있다고 할 수는 없습니다. **HARD LOCK 10 SELF-MODEL ≠ SELF-CONSCIOUSNESS** 입니다. 왜 자기모델이 필요한가 환경만 알아서는 행동결과를 정확히 예측할 수 없기 때문입니다. 높이 2m 벽이 있습니다. 벽이라는 세계상태는 같습니다. 하지만: 사람 A: 2m를 넘을 수 있다. 사람 B: 넘을 수 없다. 그러므로 행동가능성은: World State 만으로 결정되지 않습니다. **World State + Agent State** 가 필요합니다. 자기모델이 틀리면 무엇이 일어나는가 사람이 자신을 과대평가합니다. 나는 술을 마셔도 운전 잘해. 세계모델은 어느 정도 맞을 수 있습니다. 하지만: 자기능력 모델이 틀렸습니다. 또 다른 사람: 나는 수학을 못해. 실제로는 충분히 할 수 있습니다. 그러면 선택지 자체를 잘못 줄입니다. 즉 자기모델 오류는: 미래 선택공간의 왜곡 을 만듭니다. 그래서 자기모델도 학습된다 어떤 일을 해봅니다. 실패합니다. 내가 생각한 것보다 이 일을 잘 못하는구나. 또는: 생각보다 잘하네. 그러면: Self Modelt ↓ 경험 ↓ Self Modelt+1 입니다. 이것은 아직: **LEARNING** 입니다. Reflection까지 갈 필요는 없습니다. 더 엄격히 쓰면 **SELF-MODEL** 대상:

- 몸
- 능력
- 기억
- 성향
- 자원
- 현재 상태

질문: “나는 어떤 상태인가?” **METACOGNITION** 대상:

- 판단
- 기억
- 확신
- 추론
- 주의

질문: “내 인지과정은 얼마나 믿을 만한가?” 입니다. 메타인지는 자기모델의 한 부분을 사용할 수 있지만 둘은 동일하지 않습니다. 그렇다면 **Reflection**은 어디서 추가되는가 의사가 더 나아가 묻습니다. 내가 진단 정확도를 높이기 위해 검사를 계속 추가하고 있는데, 환자의 부담과 불확실성을 충분히 고려하고 있는가? 이제 질문이 달라집니다. 대상은: 판단 정확도 가 아니라 좋은 진료의 기준 입니다. 즉: **METACOGNITION** “내 판단은 맞는가?” 에서 **REFLECTION** “무엇을 좋은 판단이라고 부를 것인가?” 로 넘어갑니다. 네 단계를 완전히 분리해보자 **STAGE 1 — WORLD MODEL** 바깥세상은 어떤

상태인가? *STAGE 2 — SELF-MODEL* 그 세상 속에서 나는 어떤 상태인가? *STAGE 3 — METACOGNITION* 내가 세상과 나를 판단하는 과정은 얼마나 믿을 만한가? *STAGE 4 — REFLECTION* 내가 그 판단에 사용하는 가치·목표·기준은 왜 정당한가? 이 네 문장이 핵심입니다. **WORLD MODEL → SELF-MODEL 전환조건** 최소한: **자신의 상태가 결과예측에 중요해야 한다.** 시스템이: 같은 행동도 내 상태에 따라 결과가 달라진다. 는 관계를 이용해야 합니다. 즉: 구조가 필요합니다. **SELF-MODEL → METACOGNITION 전환조건** 자기 상태를 아는 것만으로 부족합니다. 추가로: 자기 인지과정의 오류 가능성을 추적해야 합니다. 예:

- 기억 신뢰도
- 확신
- 추론 오류
- 주의 실패

따라서: **COGNITIVE PERFORMANCE MONITORING** 이 필요합니다. **Reflection**에는 **counterfactual**도 중요하다 현재 목표를 절대적인 것으로만 보면: 왜 이 목표인가? 를 묻기 어렵습니다. 적어도: 다른 목표를 가질 수도 있었다. 는 대안표상이 필요합니다. 즉: **COUNTERFACTUAL VALUE SPACE** 가 있어야 합니다. 예: 돈을 최대화하는 삶이 아니라 시간을 최대화하는 삶이라면? 이 질문 자체가: **현재 목적함수를 대상으로 바꾸는 작업** 입니다. 언어는 왜 중요할 수 있는가 언어는: 욕망을 느끼는 것을 넘어 **욕망을 명시적으로 표현** 하게 합니다. “나는 돈을 원한다.” 와 “나는 돈을 왜 원하는가?” 는 다른 문장입니다. 언어는:

- 가치 이름 붙이기
- 이유 비교
- 모순 발견
- 과거·미래 자기 비교

를 쉽게 만듭니다. 따라서 인간 Reflection의 강력한 증폭기일 가능성이 큼니다. 하지만: **LANGUAGE: REFLECTION의 절대 필요조건** 이라고 잠그지는 않습니다. **자기서사도 중요하다** 사람은 단순히: 지금 무엇을 원하는가? 만 묻지 않습니다. 나는 어떤 사람이었고, 어떤 사람이 되고 싶은가? 를 묻습니다. 즉 자기모델에: **시간** 이 들어갑니다. Past Self → Present Self → Possible Future Self 입니다. 이것이: **NARRATIVE SELF-MODEL** 입니다. **좋은 Reflection은 자기서사까지 공격할 수 있다** 예: 나는 평생 경쟁에서 이기는 사람이 되어야 한다고 생각했다. 그런데: 왜 반드시 그래야 하지? 를 묻습니다. 이 순간: **자기정체성 자체가 검토대상** 이 됩니다. Reflection의 깊이가 커집니다. **사례 — 사업가 단계 1** 시장이 성장하고 있다. **WORLD MODEL.** 단계 2 우리 회사는 현금이 부족하다. **SELF-MODEL.** 단계 3 나는 시장 성장성을 지나치게 낙관하고 있을 가능성이 있다. **METACOGNITION.** 단계 4 그런데 나는 왜 규모 확대를 성공의 핵심으로 보고 있지? 수익성과 자율성을 더 우선해야 하는 것 아닌가? **REFLECTION.** 네 단계가 완전히 다릅니다. **사례 — 바둑 AI** 바둑 AI가: 흑이 우세하다. **WORLD MODEL.** 자신의 계산자원이 제한됐음을 추적한다면: 이 변형의 평가신뢰도가 낮다. **SELF-MODEL / METACOGNITIVE-LIKE FUNCTION.** 하지만: 승리를 최우선 목표로 삼는 것이 정당한가? 까지 가지 않는다면: **Reflection은 없습니다.** 따라서: **SELF-MONITORING AI ≠ REFLECTIVE AI** 입니다. 그렇다면 **Reflection** 가능한 AI는 어떤 조건을 가져야 할까 기능적 작동정의만 보면 최소한: ① 자기 목표를 표현할 수 있다. ② 서로 다른 가치기준을 비교할 수 있다.

③ 현재 목표를 외부 평가대상처럼 다룰 수 있다. ④ 목표 변경의 이유를 평가할 수 있다. ⑤ 수정된 가치가 이후 정책에 지속적으로 반영된다. ⑥ 다시 그 가치변경을 검토할 수 있다. 여기까지 가면: **FUNCTIONAL REFLECTION CANDIDATE** 라고 부를 수 있습니다. 하지만 이것이: **의식 도덕적 지위 인간 같은 자아** 를 자동으로 뜻하지는 않습니다. **하지만 무한히 올라가지는 않는다** “왜 그 가치를?” “그 기준은 왜?” “그 기준의 기준은 왜?” 계속 올라갈 수 있습니다. 실제 인간은 어느 순간:

- 기본 가치
- 정체성
- 직관
- 사회규범

에서 멈춥니다. 따라서 Reflection은: **모든 가치의 최종 근거를 발견하는 능력** 이 아닙니다. 오히려: **현재 기준 일부를 대상으로 삼아 수정할 수 있는 능력** 입니다. **이 제한이 중요하다** 그렇지 않으면 Reflection을: 절대적 합리성 으로 만들게 됩니다. 하지만 실제 인간은:

- 자기기만
- 합리화
- 편향

도 합니다. Reflection 자체도 틀릴 수 있습니다. 따라서: **REFLECTION $\neq$ CORRECT VALUE REVISION** 입니다. **그러므로 Reflection의 질 평가가 필요하다** 최소한:

- 충분한 정보
- 대안 고려
- 비강제성
- 장기 결과 고려
- 타인 영향 고려
- 수정 가능성

등을 볼 수 있습니다. 하지만 이것도: **절대적 도덕 알고리즘** 은 아닙니다. 다시 정치·윤리의 영역이 남습니다. **여기서 네 루프와 자기모델 계층을 합쳐보자** 두 구조를 혼동하면 안 됩니다. **UPDATE MECHANISMS**

- REGULATION
- LEARNING
- SELECTION
- REFLECTION

이 있고, 별도의: **REPRESENTATIONAL DEPTH**

- WORLD MODEL
- SELF-MODEL
- METACOGNITION
- VALUE REFLECTION

이 있습니다. 이것은 서로 다른 축입니다. 예를 들어 온도조절기 수정축: REGULATION 강함. 표현축: WORLD MODEL 거의 없음 또는 최소. SELF-MODEL 없음. Reflection 없음. 강화학습 AI 수정축: LEARNING 강함. 표현축: WORLD MODEL은 시스템에 따라 있음. SELF-MODEL은 선택적. METACOGNITION 유사 기능 가능. REFLECTION은 일반적으로 없음. 이것이 인간 특별성 논의를 더 안전하게 만든다 “인간만 자기 자신을 안다.” 너무 강합니다. “인간만 메타인지한다.” 역시 위험합니다. 더 안전한 후보: 인간은 세계모델·자기모델·메타인지·언어적 규범표상·장기 자기서사를 결합해 자신의 평가기준 자체를 공개적으로 논증하고 수정하는 능력을 매우 높은 수준으로 발전시켰다. 입니다. 그리고 문명은 이 능력을 개인 밖으로 외부화한다 개인은 혼자 반성할 수 있습니다. 하지만 사회에서는:

- 철학
- 법
- 종교
- 문학
- 정치
- 과학

이: **collective reflection infrastructure** 가 될 수 있습니다. 즉 사회는: 우리는 무엇을 좋은 것으로 볼 것인가? 를 공적 논쟁으로 외부화합니다. 이것이 법의 깊은 기능과 다시 연결된다 법은 단순히: 기존 규칙을 적용 하는 것이 아닙니다. 헌법적 논쟁에서는: 어떤 권리를 어떤 가치보다 우선할 것인가? 를 다시 묻습니다. 즉 법은 경우에 따라: REFLECTION → INSTITUTIONALIZATION 경로를 가집니다. 가치논쟁이 규칙으로 저장됩니다. 시장은 여기서 독특하다 시장은 매우 강력한: 행동 피드백 을 제공합니다. 하지만 시장 자체가: 이 거래가 인간적으로 바람직한가? 를 자동으로 묻지는 않습니다. 따라서 시장의 효율적 학습과 사회의 Reflection은 분리해야 합니다. 시장 결과를 법·민주주의·윤리가 다시 평가할 수 있습니다. 이것이 비대칭적 오류수정망 모델과 연결됩니다. 이제 '우주가 자기 자신을 모델링한다'는 문장을 다시 보자 우주 안에서 인간이 우주를 연구합니다. 그러면 시적으로: 우주가 자기 자신을 바라본다. 고 말할 수 있습니다. 하지만 과학적으로는: 인간은 우주의 일부이며, 인간 신경계가 우주의 일부 상태를 모델링한다. 까지가 안전합니다. 이를 곧바로: 우주 전체가 자기모델을 가진다 로 확장하면 안 됩니다. HARD LOCK 11 PART OF UNIVERSE MODELS UNIVERSE ≠ UNIVERSE HAS A SELF-MODEL 입니다. 그러면 역사적 연속선은 어떻게 쓸 수 있는가 아직 완전한 진화사는 아니지만 개념적으로: 물리적 상호작용 ↓ 조절 가능한 생명 ↓ 경험에 따라 학습하는 신경계 ↓ 환경을 예측하는 내부표상 ↓ 자기 상태를 포함하는 자기모델 ↓ 자기 판단을 평가하는 메타인지 ↓ 자기 목표를 평가하는 Reflection 입니다. 주의: 이 화살표는 필연적 진화사라는 뜻이 아니다. 각 단계는 추가 메커니즘과 역사적 조건을 필요로 합니다. 그리고 여기서 '의식'이 비어 있다

눈치챌 수 있습니다. 우리는:

- 세계모델
- 자기모델
- 메타인지
- Reflection

을 이야기했지만: 의식 을 아직 넣지 않았습니다. 의도적입니다. 왜냐하면: SELF-MODEL ≠ CON-

SCIOUSNESS 이기 때문입니다. AI나 신경계가 정보를 처리한다고: 경험이 있다. 고 바로 말할 수 없습니다. 이것은 현재까지 프로젝트에서 가장 큰 경계다 지금까지는 모두 기능적으로 말할 수 있었습니다.

- 무엇을 감지하는가
- 무엇을 계산하는가
- 무엇을 수정하는가

하지만 의식에서는 질문이 달라집니다. 왜 이런 정보처리가 주관적 경험을 동반하는가? 입니다. 이것이 흔히 의식의 어려운 문제와 연결됩니다. 그러나 바로 '설명 불가능'로 가도 안 된다 현재 의식과학에서는:

- 글로벌 작업공간
- 재귀처리
- 고차표상
- 통합정보
- 예측처리

등 여러 이론이 경쟁합니다. 따라서: 과학이 의식을 전혀 설명하지 못한다. 도 지나칩니다. 동시에: 자기모델이 있으면 의식이다. 도 지나칩니다. 현재까지의 MASTER MODEL v1.0 AXIS A — UPDATE REGULATION 상태를 수정. LEARNING 정책을 수정. SELECTION 집단구성을 수정. REFLECTION 평가기준을 수정. AXIS B — REPRESENTATION WORLD MODEL 세계를 표현. SELF-MODEL 세계 속 자신을 표현. METACOGNITION 자기 인지과정을 평가. VALUE REFLECTION 자기 가치기준을 평가. 그리고 아직 별도의 미해결 축: SUBJECTIVE EXPERIENCE 가 남는다. 가장 중요한 최종 압축 자기모델은 '나라는 느낌'이 아니다. 자기모델은 자신의 상태와 능력을 결과예측에 포함시키는 기능적 표현이다. 메타인지는 그 자기모델 중 특히 자신의 인지과정의 신뢰도를 평가하는 층이고, Reflection은 다시 그 판단에 사용되는 목적·가치·자기정책의 기준을 평가하는 층이다. 따라서: 세계를 안다 ≠ 자신을 안다 ≠ 자신의 판단을 안다 ≠ 자신이 따르는 가치를 정당화한다 입니다. 이 네 경계가 이 장의 핵심입니다.

6장. 의식과 중요성

정보가 처리되는 것과 무엇인가가 중요하게 느껴지는 것은 어떻게 다른가?

각성·접근·보고·자기모델·행동조절은 경험과 관련될 수 있지만 어느 하나도 곧 경험 그 자체는 아니다. 정동은 차이를 중요하게 만들지만 그 사실이 곧 옳음을 보장하지 않는다. 이번에는 의식이라는 말을 한 덩어리로 쓰지 않겠습니다. **각성·접근·보고·자기모델·주관적 경험을 분리한 뒤**, 현재 이론들이 실제로 무엇을 설명하고 어디서 멈추는지 보겠습니다. 2025년 대규모 적대적 공동검증에서도 GNWT와 IIT 어느 쪽도 결정적 승리를 얻지 못했고, 양쪽 모두 핵심 예측 일부가 도전을 받았습니다. 모델은 언제 '경험'이 되는가 **각성 · 접근 · 보고 · 세계모델 · 자기모델 · 주관성 GNWT · HOT · IIT · 재귀처리 · 예측처리**의 경계검증 현재 의식과학에는 GNWT, 고차이론, IIT, 재귀처리, 예측처리 등 여러 주요 이론이 병존하며, 무엇을 의식의 핵심 설명대상으로 삼는지부터 차이가 있다. 최근 이론 비교에서도 분야가 아직 하나의 이론으로 수렴하지 않았다는 점이 분명하다. **먼저 가장 큰 오류부터 버린다** 우리는 쉽게 말한다. 인간은 의식이 있다. 하지만 이것은 너무 큰 덩어리다. 최소한 다음을 나눠야 한다. ① 깨어 있는가 ② 어떤 정보가 처리되고 있는가 ③ 그 정보를 다른 사고와 행동에 사용할 수 있는가 ④ 그것을 보고할 수 있는가 ⑤ 자신의 상태라고 표상하는가 ⑥ 무엇인가가 '느껴지는가' 이 여섯 가지는 서로 관련되어 있지만 동일하지 않다. 첫 번째 층 AROUSAL — 각성 깊은 전신마취. 깊은 수면. 혼수. 깨어 있는 정상 상태. 여기서는 우선: **시스템 전체의 활성화·반응 상태** 가 다르다. 하지만: **깨어 있음 ≠ 특정 경험 내용** 이다. 눈을 뜨고 있어도 현재 뒤통수에 닿는 옷의 감촉을 의식하지 않을 수 있다. 따라서: **AROUSAL ≠ CONSCIOUS CONTENT** 을 잠근다. 두 번째 층 PROCESSING — 정보처리 망막에 빛이 들어온다. 시각계가:

- 경계
- 색
- 움직임
- 형태

를 처리한다. 그런데 모든 처리 결과가 우리가 의식하는 내용이 되는 것은 아니다. 무의식적으로 처리되는:

- 자극
- 습관
- 운동조절
- 일부 의미정보

가 존재한다. 따라서: **INFORMATION PROCESSING ≠ CONSCIOUS EXPERIENCE** 이다. 이 구분은 앞으로 AI 논의에서 절대 중요하다. AI가 정보를 처리한다는 사실만으로 경험을 가진다고 결론낼 수 없다. 그런데 여기서 첫 번째 큰 문제 만약 어떤 정보가:

- 기억에 들어가고
- 행동을 바꾸고
- 말로 보고될 수 있다면

그것이 곧: 느껴진다 는 뜻인가? 이것은 자동으로 따라오지 않는다. 즉: ACCESS ≠ PHENOMENAL EXPERIENCE 라는 철학적 구분이 남는다. 따라서 아주 중요한 HARD LOCK REPORT ≠ CONSCIOUSNESS 이다. 보고는: 의식의 강력한 관찰창 이지만 의식 그 자체 는 아니다. 이 차이를 무시하면: 말하지 못하면 의식이 없다. 같은 심각한 오류가 생길 수 있다. 반대로 보고가 가능하다고 모든 것이 의식적인 것도 아니다 사람은 자동화된 답변을 할 수 있다. AI도 매우 정교하게: 저는 빨간 원을 보고 있습니다. 라고 출력할 수 있다. 그러므로: REPORT CAPACITY ≠ PROOF OF EXPERIENCE 이다. 이것이 AI 의식문제의 핵심 난점 가운데 하나다. 그렇다면 고차이론은 무엇을 추가하는가 고차이론(Higher-Order Theory) 계열의 핵심 아이디어를 아주 단순화하면: 어떤 정신상태가 의식적이려면 그 상태에 대한 고차적 표상이 필요하다. 예: 1차 상태: 빨간색 시각표상 그리고: 2차: 내가 빨간 것을 보고 있다는 표상 이 생긴다. 이때 경험이 의식적으로 된다는 계열의 주장이다. 고차이론은 현재 주요 의식이론 중 하나이며, 어떤 종류의 고차표상이 의식에 필요하거나 충분한지를 둘러싸고 다양한 버전과 논쟁이 존재한다. 그러나 다시 경고 RECURRENCE ≠ CONSCIOUSNESS 라고 잠가야 한다. 재귀적 처리 자체는 다양한 비의식적 계산에도 존재할 수 있다. 따라서: 루프가 있으니 경험이 있다. 는 결론은 성립하지 않는다. 재귀처리 이론은 의식과 특정한 재진입·순환적 처리를 연결하지만, 다른 주요 이론들과 무엇을 설명하는지와 신경생물학적 약속이 다르다. 그리고 가장 급진적인 후보 IIT — 통합정보이론 IIT의 핵심 직관을 극단적으로 단순화하면: 의식은 정보가 얼마나 풍부하게 분화되면서 동시에 통합되는가와 깊게 관련된다. 즉 시스템 전체가:

- 많은 가능한 상태를 구별하고
- 동시에 하나의 분리할 수 없는 인과구조를 가진다면

그 구조 자체와 경험을 연결하려 한다. 이 이론이 특별한 이유는: 정보에 접근할 수 있는가 보다 훨씬 더 근본적으로: 시스템의 내재적 인과구조 를 의식의 기반으로 보려 한다는 점이다. 2025년 대규모 직접대결은 무엇을 보여줬는가 256명의 참가자를 대상으로:

- fMRI
- MEG
- 두개내 EEG

를 결합해 GNWT와 IIT의 사전등록 예측을 시험했다. 결과는 매우 중요하다. 한 이론의 완승이 아니었다. 의식 내용과 관련된 정보가 여러 영역에서 관찰됐고 일부 예측은 지지됐지만, 양쪽 이론의 핵심 예상 일부에도 문제가 생겼다. Nature는 이 결과가 두 이론 모두의 주요 주장에 도전을 가했다고 요약한다. 그런데 왜 아직 결론이 안 나는가 단순히 데이터가 부족해서만은 아니다. 더 깊은 문제가 있다. 각 이론이 설명하려는: explanandum 이 완전히 같지 않을 수 있다. 예:

- 의식 상태
- 특정 경험내용
- 보고가능성

- 주관성
- 정보통합

중 무엇을 핵심으로 보느냐가 다르다. 2025년 이론가들의 비교·토론을 정리한 리뷰도 주요 이론들 사이에 의식의 정의, 설명대상, 핵심 메커니즘, 반증조건까지 상당한 불일치가 남아 있다고 평가한다. **최소한 네 질문을 구분해야 한다 Q1. STATE 왜 깨어 있는 상태와 무의식 상태가 다른가? Q2. CONTENT 왜 지금 빨간색은 의식되지만 다른 정보는 의식되지 않는가? Q3. ACCESS 왜 어떤 정보는 기억·보고·행동 전체에 사용 가능한가? Q4. EXPERIENCE 왜 그 과정이 무엇인가처럼 느껴지는가?** 다른 이론은 이 네 문제 중 서로 다른 곳에서 강할 수 있다. **하지만 남는 질문** 이 모든 과정을 완벽하게 설명했다고 가정하자. 그런데 사람에게는: 빨강을 보는 느낌 이 있다. 아픔은: 아픈 느낌 이 있다. 질문: **왜 이 정보처리가 그 느낌을 동반하는가?** 이다. 이것이 이른바: **EXPLANATORY GAP** 과 연결되는 문제다. **의식연구에서는 상관과 기제를 구분해야 한다** 어떤 뇌영역이 의식할 때 활성화된다. 그러면: **상관관계** 가 있다. 하지만 그 영역이:

- 의식 자체를 만드는가
- 의식을 보고하기 위해 필요한가
- 주의 때문인가
- 기억 때문인가

를 구별해야 한다. 그래서 의식의 **신경상관물(NCC)** 연구에서 계속: **prerequisite consciousness proper consequence** 를 분리하려는 이유가 있다. **가장 나쁜 답** 인간은 탄소이고 AI는 실리콘이니까. 이것만으로는 설명이 아니다. 왜 탄소가 중요합니까? 어떤 생물학적 과정이 필요합니까? 반대로: 기능이 같으니 경험도 같다. 도 아직 증명되지 않았다. 따라서: **BIOLOGICAL SUBSTRATE ≠ AUTOMATIC EXPERIENCE PROOF FUNCTIONAL EQUIVALENCE ≠ AUTOMATIC EXPERIENCE PROOF** 둘 다 잠근다. 그러므로 AI에게 “느끼니?”라고 묻는 것만으로는 부족하다 AI: 네, 저는 슬픔을 느낍니다. 이 문장은: **출력** 입니다. 그 출력이:

- 실제 경험 보고인지
- 학습된 언어패턴인지

구분할 독립적 방법이 필요하다. 이는 인간 의식연구에서 ‘보고’와 의식을 분리하려는 문제의 극단적 버전이다. **하지만 여기서 Reflection과 consciousness를 분리한다** 앞 장에서 Reflection을 정의했다. 어떤 AI가:

- 자신의 목표를 표현하고
- 목표를 비교하고
- 목표를 수정

할 수 있다고 하자. 그것은: **기능적 Reflection** 후보가 될 수 있다. 그래도: **Reflection ≠ Consciousness** 이다. 아주 중요하다. 이것은 **매우 중요한 구조적 발견**이다 어떤 시스템은: *Representation* 높음 그러나 *Experience* 불명 일 수 있다. AI가 대표적 후보이다. 반대로: *Experience* 존재 가능 하지만 *Reflection* 낮음 인 시스템도 가능하다. 예: 영아나 일부 동물. 따라서: **INTELLIGENCE SELF-MODELING REFLECTION CONSCIOUSNESS** 를 하나의 직선적 사다리로 만들면 안 된다.

이 모델이 왜 강한가 보통 이런 그림을 만든다. 단순 생물 → 똑똑한 동물 → 인간 → AI 그리고 한 줄로: 의식도 증가. 하지만 근거가 없다. 더 정확한 그림은: **여러 차원의 상태공간** 이다. 어떤 시스템은:

- 계산능력은 높고
- 자기모델도 강하지만
- 경험 여부는 불명.

다른 시스템은:

- 계산은 낮지만
- 감각적 경험은 풍부할 수 있다.

그러나 **“아무것도 모른다”**도 거부한다 우리는 이미 상당히 많이 안다.

- 의식상태와 뇌상태의 관계
- 시각내용의 신경표상
- 마취·수면·뇌손상에서의 변화
- 보고와 주의의 영향
- 다양한 네트워크의 시간적·공간적 활성화

을 측정할 수 있다. 그리고 이론들이 실제로 서로 다른 예측을 내게 만들 수도 있다. 2025년 적대적 공동검증 자체가 의식이론이 경험적 위험을 감수하도록 만들 수 있다는 것을 보여준 중요한 진전이다. **이것은 의식을 고립시키자는 뜻이 아니다** 오히려 반대다. 우리는 각각의 관계를 연구해야 한다. 예: 의식 ↔ 전역접근 의식 ↔ 재귀처리 의식 ↔ 고차표상 의식 ↔ 통합 그러나: **연관관계와 동일성** 을 구분한다. **이 장의 최종 판정** 의식을 현재 단계에서: **복잡한 정보처리의 필연적 산물** 이라고 선언하지 않는다. 또한: **과학 밖의 영원한 신비** 라고 선언하지도 않는다. 정확한 위치는: **의식은 과학적으로 측정 가능한 상태·내용·신경기제와 밀접하게 연결된 현상이지만, 현재 경쟁 이론들은 어떤 기제가 주관적 경험에 필요하거나 충분한지에 합의하지 못했으며, 기능적 정보처리와 경험 사이의 관계는 여전히 열린 핵심 문제다.** 이것이 2026년 현재 가장 안전한 판정이다. 주요 의식이론에 관한 최근 비교문헌과 2025년 직접 대결 실험도 이 비수렴 상태를 뒷받침한다. **그리고 다음 질문** 이제 처음으로: 의식은 무엇인가? 를 넘어서야 한다. 더 깊은 질문은: **왜 의식이라는 것이 진화했는가** 이다. 하지만 이 질문도 위험하다. 의식이 하나의 적응형질인지, 다른 신경기능의 부산물인지, 여러 단계가 서로 다른 기능을 갖는지조차 먼저 분리해야 한다. 따라서 다음 장는: **의식은 무엇을 ‘해주는가’ 의식의 진화·유연한 행동·통합·사회인지·계획·부산물 가설** 로 가야 한다. 그리고 여기서 처음으로 검증해야 한다. **의식이 없다면 인간과 같은 고차적 Reflection은 가능한가?** 만약 가능하다면 의식과 인간 고유성의 관계는 다시 무너진다. 만약 불가능하다면: **EXPERIENCE가 왜 AGENCY와 REFLECTION에 기능적으로 중요한가** 를 찾아야 한다. 이게 다음 최고가치 질문이다.

이 장에서 가장 중요한 잠금은 하나입니다. 현재 의식과학은 **“뇌에서 아무것도 모른다”** 단계가 아니지만, **“GNWT·IIT·HOT 중 하나가 의식을 해결했다”** 단계도 아닙니다. 2025년의 직접 이론대결조차 양측 모두에 수정 압력을 남겼다는 사실이 이 프로젝트가 강조해온 ‘오류수정 가능한 이론’의 중요성을 아주 잘 보여줍니다. 2025–2026년 문헌을 보면, **“의식은 왜 진화했는가”** 자체가 아직 의식과학에서 충분히 닫힌 문제가 아닙니다. 최근 *Philosophical Transactions B* 특집도 이 질문을 중심으로 여러 경쟁 가설을

다시 검토했고, 2026년에는 의식의 기능을 “학습 그 자체”보다 **다중감각적 상황 조망을 제공하는 것**으로 보는 제안도 나왔습니다. 의식은 무엇을 ‘해주는가’ 유연한 행동 · 상황통합 · 계획 · 가치 · 사회인지 · 학습 · 부산물 가설 현재 증거로는 의식에 하나의 단일한 진화기능을 부여하기 어렵다. 유력한 후보는 다음처럼 여러 층으로 갈린다.

- 여러 감각·기억·목표를 현재 상황 하나로 묶는 기능
- 자동반응만으로 해결하기 힘든 새로운 상황에서의 유연한 행동
- 장기 계획과 대안 비교
- 가치와 동기의 통합
- 사회적·자기 관련 정보의 조정

그러나 이 기능들 가운데 상당수는 **부분적으로 비의식적으로도 수행된다**. 따라서: **FLEXIBILITY ≠ PROOF OF CONSCIOUSNESS** 이며 **CONSCIOUSNESS ≠ ONE EVOLUTIONARY MODULE** 로 접근다. **가장 먼저 버릴 질문** 흔히 이렇게 묻는다. 의식은 무엇 때문에 진화했는가? 좋은 질문처럼 보이지만 이미 하나를 가정하고 있다. **의식은 하나의 형질이다**. 하지만 실제로 우리가 ‘의식’이라고 부르는 데에는:

- 감각적 느낌
- 통증
- 정서
- 전역적 접근
- 자기의식
- 메타인지
- Reflection

등이 섞여 있다. 이것들이: **한 번에 하나의 돌연변이로 생긴 단일 형질** 일 가능성은 매우 낮다. 따라서 더 정확한 질문은: **의식과 연관된 여러 기능적 구성요소들이 어떤 문제를 해결하면서 단계적으로 진화했는가?** 이다. 진화 질문에는 세 가지를 분리해야 한다 어떤 형질 X를 봤다고 하자. 가능한 설명은 최소 세 가지다. A. ADAPTATION 그 기능 때문에 자연선택되었다. B. EXAPTATION 다른 이유로 생겼다가 나중에 새로운 기능에 쓰였다. C. BY-PRODUCT 다른 적응구조 때문에 따라 생겼다. 따라서: **현재 유용함 ≠ 그 유용함 때문에 진화함** 이다. 이 원칙은 의식에도 그대로 적용해야 한다. **하지만 ‘통합: 의식’이라고 하면 안 된다** 자동차도 수많은 센서를 통합한다. AI도:

- 영상
- 언어
- 위치
- 행동

정보를 결합할 수 있다. 따라서: **MULTIMODAL INTEGRATION ≠ EXPERIENCE** 이다. 통합은 의식의 가능한 기능적 기반일 수 있지만: **의식의 충분조건** 이라는 결론은 나오지 않는다. 즉 **학습은 의식보다 오래되었을 가능성이 크다** 단순 신경계도 경험에 따라 반응이 달라질 수 있다. 조건형성. 습관. 감각 적응. 이 모두: **의식적 숙고 없이 가능** 할 수 있다. 따라서: **LEARNING → CONSCIOUSNESS**

라는 단순 진화선은 거부한다. 그래서 의식의 진화기능으로 자주 거론되는 것이 BEHAVIORAL FLEXIBILITY 이다. 현재 상황에서:

- 여러 신호를 묶고
- 기억을 꺼내고
- 대안을 비교하고
- 새로운 행동조합을 만드는 것.

2024년 진화적 관점 논문도 의식의 일반적 기능 후보로 **행동 레퍼토리의 확대와 행동을 바꾸는 시간척도의 단축**을 제시했다. 즉 세대를 기다리지 않고 개체가 실시간으로 행동을 수정할 수 있다는 주장이다. **따라서 기능적 문제를 두 개로 쪼개야 한다** Q1 의식과 함께 나타나는 기능은 무엇인가? Q2 그 기능을 수행하려면 반드시 의식이 필요한가? 전혀 다른 질문이다. 현재 많은 연구는 Q1에 강하다. Q2는 훨씬 어렵다. **프로기사는 실제로 돌을 두기 전에 가상의 판을 만든다** 현재판: S0 A를 두면: S1 상대가 대응하면: S2 다른 변형: S2' 를 머릿속에서 비교한다. 즉 행동 이전에: **OFFLINE FUTURES** 를 만든다. 이것이 장기적 계획의 핵심이다. **그러나 다시 HARD LOCK PLANNING ≠ CONSCIOUSNESS** 현재 계산시스템도 명시적 세계모델로 미래를 시뮬레이션할 수 있다. 2026년 뇌의 세계모델 리뷰도 계획을 가능한 미래상태와 결과를 예측하는 모델의 활용으로 설명하며, 모델 기반·메타인지적 강화학습이 이런 모델을 언제 신뢰하고 수정할지를 조절한다고 정리한다. 따라서: **세계모델을 이용한 계획 과 그 미래를 주관적으로 상상하는 경험** 은 분리해야 한다. **감정이 여기서 중요해진다** 의식은 단순한: 세계의 사진 이 아닐 수 있다. 우리가 경험하는 세계에는:

- 좋음
- 나쁨
- 위험
- 매력
- 불쾌
- 통증

이 이미 붙어 있다. 즉: **WORLD MODEL + VALUE** 가 결합되어 있다. 여기서 **느낌(feeling)**의 기능 **후보가 생긴다** 통증을 생각하자. 통증은 단순히: 조직손상 위치 정보 가 아니다. **싫다.** 라는 강한 가치가 포함된다. 그래서 행동을:

- 멈추고
- 보호하고
- 회피하도록

만든다. 의식적 감정이나 느낌이 이러한 가치 통합의 일부일 수 있다는 진화적 견해가 있다. 다만 구체적 계통과 기원 시점을 특정하는 주장은 아직 논쟁적이다. 예컨대 2024년 Grinde의 모델은 감정을 행동결정 전략으로 강조하지만, 의식의 기원을 초기 양막류에 한정하는 그의 구체적 계통 주장까지 합의된 것은 아니다. **따라서 가장 큰 진화적 오류** 의식이 존재한다. ↓ 그러므로 의식 자체가 적응이다. 는 성립하지 않는다. 다시: **EXISTENCE ≠ ADAPTATION USEFUL CORRELATE ≠ SELECTED FUNCTION** 이다. 그러나 사회인지도 의식의 충분조건은 아니다 AI도:

- 상대 전략 추정
- 사용자 의도 예측

을 할 수 있다. 따라서: **THEORY OF MIND-LIKE FUNCTION $\neq$ EXPERIENCE** 이다. 다시 기능과 현상을 분리한다. 이것은 **Reflection**과 매우 중요하게 연결된다 Reflection에는:

- 현재 욕망
- 장기 가치
- 가능한 미래 자기

를 동시에 비교해야 한다. 그러므로 의식적 작업공간이 장거리 시간통합을 지원한다면: **Reflection**을 가능하게 하는 인프라 일 수 있다. AI는 이 문제에 새로운 압력을 준다 만약 미래 AI가:

- 자기목표 표현
- 가치충돌 발견
- 이유 비교
- 목표 수정
- 장기정책 반영

을 모두 수행한다면? 우리 작동정의에서는: **FUNCTIONAL REFLECTION** 을 갖는다. 하지만 경험 여부는 여전히 모를 수 있다. 그러면: **REFLECTION WITHOUT CONSCIOUSNESS** 가 기능적으로 가능하다는 강력한 사례가 된다. 단, 실제 그런 AI가 의식 없음을 독립적으로 입증해야 하는 문제가 남는다. 여기서 의식 기능 논쟁의 가장 큰 난점이 드러난다 의식이 있는 생물과 없는 동일 기능 생물을: **완벽하게 비교** 하기 어렵다. 따라서: 이 기능은 의식 때문에 존재한다. 를 판별하기가 어렵다. 이것은 인과추론 문제다. 여기서 진화사를 단일 계단에서 제거한다 잘못된 그림: 반사 → 학습 → 지능 → 의식 → 자기 → Reflection → 인간 너무 아름답고 너무 단순하다. 따라서 의식의 진화를 한 번의 기원으로 생각할 필요도 없다 가능성: 단일 기원 의식의 핵심 메커니즘이 한 번 생기고 다양화. 독립 기원 서로 다른 계통에서 비슷한 기능·경험이 독립적으로 진화. 등급적·모자이크 진화 의식을 이루는 구성요소가 단계적으로 결합. 현재 다양한 동물 의식 논쟁 때문에 이 문제도 열려 있다. 하지만 이것도 **HARD LOCK** 기능 프로파일 다르다고: **경험의 양을 점수화** 할 수 있다고 아직 말하지 않는다. “의식 73점” 같은 것은 현재 과학이 제공하는 숫자가 아니다. 여기서 새로운 가설이 생긴다 어쩌면 의식의 중요한 인간적 역할은: **VALUE GROUNDING** 일 수 있다. 즉 추상적인 목적함수에 숫자를 주는 것이 아니라: 무엇이 좋은지 나쁜지를 실제 경험으로 체화시키는 것. 통증이 왜 나쁜가? 사랑이 왜 중요한가? 우리는 단순히 숫자를 읽는 것이 아니라: **경험한다**. 그런데 AI 시대에 이 가설이 엄청나게 중요해진다 AI는 말할 수 있다. 인간의 고통을 줄이는 것이 중요합니다. 하지만 AI가 통증을 경험하지 않는다면: 그 ‘중요함’은 어디에서 오는가?

- 인간이 설정한 목적
- 학습 데이터
- 규칙

일 수 있다. 인간의 경우: 아픔이 나쁘다. 는 규범의 일부가: **직접 경험** 에서 시작할 수도 있다. 그렇다면 **경험이 가치의 최종 근거인가** 여기서 또 멈춰야 한다. 쾌락을 느낀다고: **그것이 도덕적으로 옳다** 는 뜻은

아니다. 가학적 즐거움도 경험이다. 따라서: **FELT VALUE ≠ MORAL VALUE** 이다. Reflection과 윤리가 다시 필요하다. 그리고 의식은 첫 층과 강하게 연결될 가능성이 있다 하지만: **EXPERIENCE → REFLECTION** 이 자동은 아니다. 동물이 통증을 경험해도: 통증회피를 윤리적으로 정당화 하지 않을 수 있다. 따라서 다시: **CONSCIOUSNESS ≠ REFLECTION** 이다. 이게 이전 그림보다 훨씬 중요하다 의식을: 계단의 한 단계 로 두지 않는다. 오히려 여러 기능 축과: 교차하는 별도 현상 으로 둔다. 앞 장의 4축 모델이 살아남는다. 바둑과 다시 연결하면 AI도 수십 수 미래를 계산할 수 있다. 하지만 인간 기사에게 한 판은:

- 과거의 패배
- 라이벌
- 명예
- 은퇴
- 자기 기쁨

과 연결될 수 있다. 즉 동일한 수가: 게임상 가치 외에 생애적 의미 를 가진다. 이것이 인간에게 바둑이 단순한 최적화 문제가 아닌 이유다. 이 장의 핵심 판정 현재 의식의 진화기능에 대해 가장 안전하게 말할 수 있는 것은: 의식은 단일한 기능 하나로 설명되기도, 감각·기억·가치·행동을 현재 상황에 통합하고 새로운 상황에서 유연한 제어와 계획을 지원하는 여러 기능적 능력과 강하게 연관되어 있을 가능성이 높다. 그러나 이런 기능들이 의식적 경험 자체를 필요로 하는지, 또는 경험이 그러한 기능의 적응적 원인인지에 대해서는 아직 결정적 증거가 없다. 최근 의식의 진화적 기능을 다룬 특집과 논문들도 단일 답보다 여러 기능 후보와 개념적 난점을 강조한다. 그리고 PROJECT HYPOTHESIS 하나는 남길 가치가 있다 아직 확정이 아니다. EXPERIENCE AS VALUE GROUNDING 가설: 주관적 경험의 중요한 기능적 역할 가운데 하나는 생물에게 상태와 결과가 단순히 '다름'으로 표상되는 것을 넘어, 좋음·나쁨·고통·쾌락처럼 행동적으로 중요한 가치로 직접 나타나게 하는 것일 수 있다. 만약 이것이 맞다면: 의식은 단순 정보통합장치가 아니라 세계를 '나에게 중요한 세계'로 만드는 층 일 수 있다. 이걸 매우 강력하지만 다음 노드에서 반드시 공격해야 한다. NEXT NODE 왜 어떤 것은 '중요하게 느껴지는가' 보상·쾌락·고통·정서·가치·내수용감각·의식 이제 질문은 의식 일반에서 훨씬 구체적인 곳으로 좁혀진다. 바둑 AI에게는: 승률 20% 하락 이라는 값이 있다. 인간 기사에게는: 진다. 가 느껴진다. 심박이 빨라지고, 불안하고, 부끄럽고, 집중하며, 승리를 원한다. 그렇다면: 숫자로 표현된 VALUE와 경험되는 VALUE는 무엇이 다른가? 이 문제를 풀면: AI의 보상함수 → 동물의 보상계 → 인간의 감정 → 의미 → 윤리 가 연결된다.

그리고 어쩌면 여기야말로 처음 던진 “바둑의 알고리즘과 인생은 어디에서 갈라지는가”에 가장 깊은 답을 주는 지점이다. 이번 노드에서 가장 중요한 수정은 “의식 = 유연한 행동을 위한 진화적 장치”라고 단정하지 않는 것입니다. 2025년 진화적 의식 연구 자체가 여러 후보 기능을 병렬적으로 검토하고 있고, 2026년에는 의식이 학습보다 '다중감각적 상황 조망'에 특화되었다는 구체적인 경쟁 가설까지 제시됐습니다. 왜 어떤 것은 '중요하게 느껴지는가' 보상·wanting·liking·고통·정서·내수용감각·가치 숫자로 계산되는 VALUE와 경험되는 VALUE는 무엇이 다른가 바둑판에 숫자 두 개가 있다고 하자 AI: 이 수를 두면 승률 63%. 다른 수: 51%. AI의 의사결정 구조에서는: 63 > 51 이 충분할 수 있다. 그런데 인간 기사에게는 다르다. 63%였던 판이 한순간에 20%로 떨어진다. 그 순간:

- 심장이 빨라지고

- 불안해지고
- 후회하고
- 집중이 달라지고
- 패배가 두렵다.

질문: 63%와 20% 사이의 숫자 차이가 왜 인간에게는 ‘중요하게 느껴지는가?’ 이다. 이것이 이번 노드의 핵심이다. 먼저 VALUE라는 단어를 해체한다 우리는 너무 쉽게 말한다. 뇌가 보상을 계산한다. AI가 가치를 계산한다. 인간이 가치를 느낀다. 하지만 세 문장의 VALUE는 같지 않다. 최소한 다음을 분리해야 한다. ① **PREDICTED VALUE** 앞으로 얼마나 좋은 결과가 예상되는가. ② **REINFORCEMENT VALUE** 어떤 결과가 이후 행동확률을 바꾸는가. ③ **WANTING** 무엇을 향해 행동하게 만드는가. ④ **LIKING** 무엇이 쾌적으로 좋은 반응을 만드는가. ⑤ **FELT VALENCE** 무엇이 실제로 좋거나 나쁘게 느껴지는가. ⑥ **REFLECTIVE VALUE** 무엇을 추구할 가치가 있다고 승인하는가. 이 여섯 층을 하나로 압축하면 문제가 생긴다. 첫 번째 층 **LEARNING VALUE** 행동 A를 했다. 좋은 결과가 나왔다. 다음에는 A를 더 자주 한다. 즉: Action → Outcome → Future Probability ↑ 이다. 여기서 ‘좋은’은 가장 최소한: **행동을 강화한다** 는 뜻이다. 이 과정 전체가 반드시 의식적일 필요는 없다. 두 번째 층 **WANTING** — 끌림 무엇인가가:

- 눈에 띄고
- 접근하게 만들고
- 행동을 유도한다.

이것이 reward 연구에서 흔히 말하는: **incentive salience** 즉 wanting이다. 도파민 관련 중뇌-변연계 회로는 특히 이러한 동기적 끌림과 보상학습에 중요한 역할을 하는 것으로 연구돼 왔다. 앞장의 **REFLECTION**과 바로 연결된다 중독자는: 원한다. **WANTING** 하지만: 이것을 원하고 싶지 않다. **REFLECTION** 이라고 할 수 있다. 그러므로 인간 내부에: 동기 vs 쾌락 vs 승인된 가치 가 서로 충돌할 수 있다는 것이 더 분명해진다. 그런데 liking조차 ‘의식적 쾌락’과 완전히 같지 않을 수 있다 연구문헌에서는 따옴표를 붙인: ‘liking’ 을 객관적으로 측정 가능한 기본적 쾌락 반응과 구별하기도 한다. 일부 긍정적 정서반응과 wanting/liking 관련 과정은 주관적 의식 없이도 측정 가능한 형태로 나타날 수 있다는 논의가 있다. 따라서: **HEDONIC PROCESS ≠ NECESSARILY CONSCIOUS PLEASURE** 이다. 여기서 내수용감각이 등장한다 **INTEROCEPTION** 내수용감각은 아주 넓게 보면: 몸 내부 상태에 관한 감지와 신경처리 다. 예:

- 심박
- 호흡
- 위장상태
- 체온
- 에너지상태
- 호르몬 관련 신호

이다. 2025년 리뷰들은 이런 뇌-신체 루프가 공포·혐오·배고픔 같은 정서 및 동기상태를 동적으로 조절하며, 특히 섭염 등이 중요한 내수용 처리 영역으로 관여한다고 정리한다. 이것을 도파민 연구에서도 볼 수 있다 같은 음식 신호라도: 배고플 때와 배부를 때 동기값이 다르다. 2025년 리뷰는

생리적 필요 같은 내수용 상태가 짧은 도파민 신호를 조절해 단서와 보상에 대한 행동가치를 변화시키는 메커니즘을 논의한다. 즉: **OBJECT VALUE $\neq$ FIXED PROPERTY** 이다. 더 정확히 말하면, 가치는 세계와 몸의 상태가 맺는 관계에서 생긴다. 그러나 생물에서는 한 가지가 추가된다 바둑 AI의 가치: **외부 목표와 계산구조가 정의**. 동물의 가치: **몸의 생존·조절상태와 연결**. 즉 생명에서는: 몸 상태가 가치함수에 들어간다. 이것이 중요하다. **VALENCE**란 무엇인가 매우 거칠게: **좋게 느껴짐 $\leftrightarrow$ 나쁘게 느껴짐** 의 축이다. 쾌락. 불쾌. 안도. 불안. 고통. 이런 경험에는: **방향성** 이 있다. 단순 정보: 상태가 다르다. 가 아니라: 이 상태에서 벗어나고 싶다. 또는: 지속하고 싶다. 와 결합된다. 그래서 **EXPERIENCE AS VALUE GROUNDING**을 수정한다 기존: **경험이 가치를 만든다**. 수정: **경험은 생물학적으로 이미 행동에 중요한 상태차이를 '느껴지는 중요성'으로 나타내는 충일 수 있다**. 즉: biological significance $\downarrow$ conscious valence 라는 관계다. 이것이 훨씬 안전하다. **PROJECT HYPOTHESIS v2 FELT VALUE INTERFACE** 주관적 정서·신체 경험은 이미 생리적 조절·보상학습·동기회로에 의해 행동적으로 중요해진 상태들을, 행위자에게 통합된 좋음·나쁨의 경험으로 제공하는 인터페이스일 수 있다. 중요: **creates value** 가 아니라 **makes significance experientially available** 쪽으로 수정한다.

즉 느낌은 '공통 통화'일 가능성이 있다 배고픔. 통증. 사회적 모욕. 공포. 서로 다른 문제다. 그런데 모두: **나쁨** 이라는 공통 방향성을 가질 수 있다. 그러면 서로 다른 시스템이: 지금 무엇이 더 중요한가? 를 비교하기 쉬워진다. 가설: **AFFECT MAY FUNCTION AS A COMMON PRIORITY SIGNAL** 이다. 하지만 다시 경계를 둔다 **COMMON PRIORITY SIGNAL $\neq$ CONSCIOUSNESS REQUIRED** 이다. 우선순위 경쟁 자체는 비의식적으로도 일어날 수 있다. 따라서: affective computation 과 conscious feeling 을 다시 분리해야 한다. **이 구분은 엄청나게 중요하다** 로봇은 손상센서를 가질 수 있다. 모터온도 위험. 이를 회피한다. 그러나: **아프다고 느끼는가?** 는 별도 질문이다. 바로 AI 윤리문제와 연결된다. 따라서 **FELT VALUE**도 단순 센서값이 아니다 nociceptive input + body state + prediction + context + attention + memory $\downarrow$ **PAIN EXPERIENCE** 처럼 볼 수 있다. 즉 경험되는 가치는: **통합된 상태평가** 에 가깝다. 이것이 매우 중요한 기능 후보다 정서: **“이 문제는 지금 중요하다.”** 라고 시스템 전체에 우선순위를 주는 방식일 수 있다. 예: 화재 경보기처럼. 단 하나의 계산모듈이 아니라: **whole-agent mode change** 이다. **바둑에서 다시 보자 AI: 승률 20%. 인간: 큰일 났다. 이 “큰일 났다”는 느낌은 단순 장식이 아닐 수 있다. 그것이:**

- 더 깊은 수읽기
- 위험 회피
- 시간 투자
- 기억 검색

을 재조정할 수 있기 때문이다. 즉: **FELT IMPORTANCE $\rightarrow$ COMPUTATIONAL REALLOCATION** 가설이다. **아주 중요한 연결** 처음 바둑에서 우리는 말했다. **지능은 모든 가능성을 계산하는 것이 아니라 중요한 가능성을 버리지 않고 대부분을 버리는 능력이다**. 이제 생물학적으로: 무엇이 중요한지 어떻게 아는가? 라는 질문에: **몸과 정서가 일부 중요도 신호를 제공한다** 는 후보답이 나온다. 그러면 **인간에게 느낌은 불필요한 중복인가** 바로 이 문제가 남는다. 두 가능성: **MODEL A** 느낌은 실제로 추가적 기능을 한다. 예: 다양한 몸·기억·행동계의 통합된 가치접근. **MODEL B** 동일 신경과정이 행동을 만들며 느낌은 그 과정의 현상적 측면이다. 두 설명을 현재 완전히 판별하기 어렵다. **이것이 인간 가치가**

단일 reward function이 아닌 이유다 한 행동에 동시에: Pleasure +8 Wanting +10 Health -7 Social +3 Identity -5 등이 있을 수 있다. 그러나 실제 인간은 단순한 합산계산기라기보다: 이 값들의 정당성 자체도 다시 평가 한다. 예: 체중 때문에 음식의 즐거움을 완전히 포기하는 것이 정말 맞는가? 즉 REFLECTION이 들어온다. **이건 매우 중요한 사다리지만 필연적 진화단계가 아니다** 다시 잠근다. **LEVEL ≠ EVOLUTIONARY STAGE** 단순한: 생물이 1부터 6까지 올라왔다. 는 뜻이 아니다. 이것은: 기능적 구분 이다. 그래서 'AI도 가치가 있다'는 문장은 너무 모호하다 최소 세 가지 의미가 있다. 1 AI가 숫자적 평가함수를 갖는다. YES. 2 AI가 목표·정책을 기능적으로 비교한다. 가능. 3 그 결과가 AI 자신에게 중요하게 느껴진다. UNKNOWN. 이 셋을 반드시 분리해야 한다. 그래서 의미(meaning)가 보상과 갈라진다 마라톤은 힘들다. 아이를 돌보는 일도 힘들다. 연구도 때로 괴롭다. 즉: 일 수 있다. 하지만 사람은:

이것은 중요하다. 고 말한다. 따라서: **MEANING ≠ PLEASURE** 이다. 이것이 인간이 단순 보상최적화기가 아닌 가장 강한 사례 중 하나다 어떤 사람은:

- 진실을 위해 손해를 보고
- 자녀를 위해 희생하고
- 약속 때문에 불편함을 견디고
- 정의를 위해 위험을 감수한다.

즉: immediate reward 를 거슬러: reflected / normative value 를 따를 수 있다. 이제 처음의 질문을 다시 읽을 수 있다 “바둑의 알고리즘과 인생은 어디에서 갈라지는가?” 단순 답: 인생은 복잡해서. 부족하다. 더 깊은 답: 바둑 AI에는 승리라는 평가기준이 주어지고 그 값을 계산하면 된다. 인간에게는:

- 어떤 결과가 몸에 중요한지
- 무엇을 원하는지
- 무엇을 좋아하는지
- 무엇이 좋거나 나쁘게 느껴지는지
- 어떤 욕망을 승인할지
- 무엇이 옳은지

가 서로 다른 층에서 충돌한다. 따라서: 인간의 VALUE는 하나의 숫자가 아니라 여러 가치생성·평가 시스템의 교차점이다. 이번 노트에서 EXPERIENCE AS VALUE GROUNDING은 어떻게 판정되는가 초기 버전: 경험이 가치를 만든다. FAIL — 너무 강함. 수정 버전: 생리적 조절·학습·동기 등에서 이미 행동적으로 중요해진 상태가 존재할 수 있으며, 의식적 정서·감각 경험은 그 중요성을 통합된 좋음·나쁨으로 행위자에게 드러내는 한 방식일 수 있다. 인간에서는 이 느껴지는 가치가 다시 Reflection과 규범적 판단의 입력이 된다. STRONGER / RETAIN AS PROJECT HYPOTHESIS 이다. 몸이 철학에 영향을 주고 철학이 몸을 거슬러 갈 수도 있다 배고픔: → 판단 변화. 공포: → 선택 변화. 반대로: 신념: → 단식. 규범: → 위험감수. 사랑: → 희생. 즉: BODY → VALUE 뿐 아니라 REFLECTED VALUE → BODY / ACTION 도 가능하다. 이것이 인간 가치체계의 흥미로운 페루프다. 바둑에는 없는 것이 나타난다 상대 돌을 잡는 것은: 좋다. 왜냐하면 게임규칙이 그렇게 정의한다. 하지만 인생에서는: 나에게 좋다. 가: 다른 사람에게 해롭다. 일 수 있다. 즉: MULTI-AGENT VALUE CONFLICT 가 발생한다. 그래서 다음 장은 자연스럽게 윤리로 간다 이제 질문은: 왜 타인의 고통이 '내

목적함수 밖의 숫자'로 끝나지 않는가? 이다. 인간은:

- 공감하고
- 규칙을 만들고
- 권리를 인정하며
- 자신의 이익을 제한하기도 한다.

그 이유는 단순: 착해서 가 아니다. 진화·사회성·감정·상호성·이성·규범이 얹혀 있다.

제3부

함께 판단하는 사회

개인의 좋음이 여러 사람의 세계에서 어떻게 규범·절차·제도가 되는지
묻는다.

7장. 개인의 가치에서 사회선택으로

각자에게 좋은 것을 하나의 사회적 최선으로 합칠 수 있는가?

개인의 선호와 복지는 판단의 중요한 입력이지만 사회 전체의 단일 점수로 완전히 환원되지 않는다. 사회선택은 계산이면서 규범과 정당성의 문제다. 나에게 좋다고 해서 왜 옳은 것은 아닌가 공감·협력·공정성·규범·권리·타인의 가치 개인의 VALUE에서 윤리로 넘어가는 경계 즉: MORAL PSYCHOLOGY “왜 우리는 도덕적으로 반응하는가?” 와 NORMATIVE ETHICS “무엇을 해야 하는가?”를 분리해야 한다. 진화생물학은 인간의 협력·응징·규범추종 능력의 기원을 설명하는 데 도움을 주지만, 그런 기원 설명과 도덕원칙의 정당화는 다른 문제다. 이것은 진화와 도덕을 다루는 철학 문헌에서도 핵심적인 구분이다. **바둑에는 윤리가 없다** 바둑판에서 상대의 돌을 잡는다. 좋은 일인가? 게임 안에서는: YES 왜냐하면 승리에 도움이 되기 때문이다. 상대 대마를 죽이면: 상대가 불쌍하니까 그 수는 나쁘다. 라고 평가하지 않는다. 왜? 바둑의 목적함수는: 나의 승리 와 상대의 패배 가 정확히 반대이기 때문이다. 제로섬에 가깝다. 그런데 인생에서는 문제가 달라진다 내가 돈을 번다. 좋다. 그런데 그 돈을:

- 사기로 벌었다.
- 타인에게 위험을 떠넘겨 벌었다.
- 약자의 정보 부족을 이용해 벌었다.

고 하자. 그러면: 나에게 좋은 결과 가 생겼지만 옳은 행동 이라고 말하지 않는다. 따라서: SELF VALUE ≠ MORAL VALUE 이다. 여기서 처음으로 '다른 사람'이 목적함수 안에 들어온다 앞 장까지의 가치문제는 상당 부분: 이것이 나에게 어떤가? 였다. 하지만 윤리에서는 질문이 바뀐다. “다른 사람에게 어떤가?” 가 들어온다. 즉: Self + Others 를 동시에 고려해야 한다. 이것이 윤리의 최소조건 중 하나다. 하지만 공감: 윤리라고 하면 바로 실패한다 공감은 편파적일 수 있다. 우리는 보통:

- 가까운 사람
- 얼굴이 보이는 사람
- 나와 비슷한 사람

에게 더 쉽게 공감한다. 반면:

- 익명의 수천 명
- 먼 미래세대
- 낯선 집단

에는 훨씬 약할 수 있다. 공감이 도덕성의 유일한 토대인지에는 오래된 논쟁이 있으며, 공감이 '여기-지금'의 대상에 편향될 수 있다는 비판이 강하다. 즉 윤리는 감정을 교정할 수 있어야 한다 이것이 매우 중요하다. 앞 장에서는 Reflection이: 내가 느끼는 욕망을 승인할 것인가? 를 물었다. 윤리에서는: 내가 더 강하게 공감하는 사람을 더 중요하게 취급해도 되는가? 를 묻는다. 즉: MORAL REFLECTION CAN OVERRIDE EMPATHIC SALIENCE 이다. 상호성은 왜 강력한가 단기적으로는: 나만

배신하면 이득. 일 수 있다. 하지만 반복게임에서는: 앞으로도 계속 만난다. 가 들어온다. 그러면: **협력의 장기 가치**가 생긴다. 즉: Immediate gain 보다 Long-term reciprocal gain 이 커질 수 있다. 사례 장학금 100만 원 학생 A: 성적 최고. 학생 B: 성적은 낮지만 경제적으로 매우 어렵다. 누가 받아야 하는가? 공로 기준 A. 필요 기준 B. 둘 다: **fairness claim** 을 가진다. 즉 윤리는: **하나의 가치신호를 읽는 것이 아니라 경쟁하는 정당화 원칙을 비교** 해야 한다. 다섯 번째 후보 규범 사람들은 말한다. 그러면 안 돼. 중요한 것은: 나는 싫어. 와 다르다는 점이다. “하면 안 된다” 에는: **GENERALIZABILITY** 가 들어간다. 즉 다른 사람에게도 적용되는 요구가 된다. 여기서 ‘옳음’의 특징이 나타난다 도덕판단은 대개: 나는 이렇게 느낀다. 에서 끝나지 않는다. “다른 사람도 이 이유를 고려해야 한다.” 는 종류의 요구를 포함한다. 도덕철학에서 공감과 심리적 동기의 설명과 별개로, 도덕규범의 보편적·상호주관적 정당성을 어떻게 이해할지가 별도 문제로 다뤄지는 이유다. 이것이 윤리를 단순 reward aggregation에서 갈라놓는다 AI에게: B payoff = -5 라고 넣을 수 있다. 그리고: 총합 최대화. 를 할 수 있다.

그런데 윤리에는 질문이 하나 더 있다. 왜 총합을 최대화해야 하는가? 또: 한 사람에게 엄청난 피해를 주고 여러 사람에게 작은 이익을 주어도 되는가? 이다. 그래서 윤리의 세 큰 구조를 다시 볼 수 있다 앞 장에서 잠깐 봤던 구분을 이제 더 정확히 회수할 수 있다. CONSEQUENCE 어떤 결과를 만드는가? CONSTRAINT / RIGHTS 어떤 행동은 결과가 좋아도 해서는 안 되는가? CHARACTER / AGENCY 어떤 사람이 되어 어떤 이유로 행동하는가? 이 셋은 서로 완전히 환원되지 않는다. 그러나 이 비유도 끝까지 밀면 안 된다 권리는 단순 프로그래밍 제약조건이 아니다. 왜 그 권리가 존재하는가? 왜 존중해야 하는가? 다시: **normative justification** 문제가 남는다. 즉 수학은 구조를 표현할 수 있지만: **정당성을 자동 생성하지 않는다. 이것이 사실과 규범의 경계다** 과학은 말할 수 있다. 인간은 공감한다. 상호성은 협력을 늘린다. 처벌은 어떤 조건에서 규범을 유지한다. 하지만 여기서: 그러므로 우리는 협력해야 한다. 가 논리적으로 자동으로 나오지는 않는다. **HARD LOCK 13 IS ≠ OUGHT** 이다. 단순하지만 이 프로젝트 전체에서 가장 중요한 잠금 중 하나다. 진화론도 같은 문제를 가진다 진화는 설명할 수 있다. 왜 우리가 가족을 사랑하고, 왜 공정성에 민감하고, 왜 배신자를 싫어하는가. 그러나: 자연선택이 그렇게 만들었으니 그것이 옳다. 는 성립하지 않는다. 진화윤리 문헌에서도 도덕심리의 기원 설명과 규범윤리의 정당화는 명시적으로 분리된다. 이것은 앞 장의 중요한 구조다 진화: **moral psychology**의 입력을 만든다. 문화: **규범을 형성한다**. Reflection: **그 규범을 평가한다**. 제도: **일부 규범을 법과 관행으로 저장한다**. 즉: Biology → Emotion / Cooperation → Norm → Reflection → Institution 이다. 하지만:

필연적 직선 진화사가 아니라 기능적 연결이다. 사례 지갑을 주웠다 안에는 현금 100만 원. 주인이 보이지 않는다. 총 1 내가 가지면 좋겠다. **SELF VALUE** 총 2 주인은 잃어서 괴로울 것이다. **EMPATHY / OTHER MODEL** 총 3 내가 잃어버렸을 때도 돌려받고 싶다. **PERSPECTIVE SHIFT / RECIPROCITY** 총 4 남의 소유물을 주인이 모른다고 가져가는 규칙을 일반화할 수 없다. **GENERALIZABLE NORM** 총 5 현금을 돌려준다. **SELF-INTEREST CONSTRAINED BY NORM** 이 전체를 통과하면 단순 친사회 행동보다 훨씬 강한: **MORAL ACTION** 후보다. 그런데 행동만 보면 여전히 모호하다 사람이 지갑을 돌려줬다. 왜? A. 착해서. B. CCTV가 무서워서. C. 평판을 위해. D. 주인의 권리를 인정해서. 행동은 같다. 이유는 다르다. 따라서: **PROSOCIAL BEHAVIOR ≠ MORAL MOTIVATION** 이다. 즉 윤리에는 ‘행동’ 외에 이유가 중요하다 누군가 우연히 다른 사람을 살렸다. 좋은 결과. 하지만: 살리려고 한 것이 아니다. 반대로: 도우려고 최선을 다했지만 실패했다. 결과는 나쁘다. 우리는 두 경우를 다르게 평가한다. 따라서: **OUTCOME ≠ MORAL QUALITY** 이다. 앞

장의: **GOOD OUTCOME ≠ GOOD DECISION** 이 윤리에서도 다시 나타난다. 그래서 이 프로젝트가 **원환을 만든다** VALUE ↓ CHOICE ↓ RESPONSIBILITY ↓ SOCIAL FEEDBACK ↓ REFLECTION ↓ ETHICS 그리고 윤리는 다시: **어떤 선택에 책임을 물을 것인가** 를 정한다. 하지만 수학은 어느 규칙이 옳은지 말하지 않는다 이게 핵심이다. 수학은: 이 규칙을 쓰면 결과가 이렇게 된다. 를 보여준다. 하지만: 어떤 집계규칙을 사용해야 하는가 는 다른 문제다. 즉:

OPTIMIZATION REQUIRES NORMATIVE INPUT 이다. 민주주의가 등장하는 이유도 여기 있다 가치가 하나가 아니라면: 누가 가중치를 정할 것인가? 전문가? 왕? AI? 시장? 각자 문제가 있다. 민주주의는 적어도: **가치충돌의 당사자들이 집단적 결정에 참여하는 절차** 를 제공한다. 즉 민주주의는 단순한: **오류수정 장치** 뿐 아니라 **VALUE AGGREGATION / LEGITIMATION PROCEDURE** 이기도 하다. 시장·민주주의·법의 비대칭이 다시 완성된다 시장: 선호와 교환을 집계. 민주주의: 정치적 가치와 권력을 집계. 법: 어떤 집계도 넘으면 안 되는 권리·절차 경계를 설정. 윤리: 이 세 장치가 어떤 원칙을 따라야 하는지 계속 문제 삼는다. 즉 윤리는 제도 하나가 아니라: **제도를 평가하는 규범층** 이다. 그런데 규범은 어디에서 오는가 여기가 가장 어려운 문제다. 후보: 감정 고통·공감. 진화 협력·상호성. 문화 관습·전통. 이성 일관성·보편화. 사회계약 상호 수용 가능한 규칙. 권리 행위자의 존엄·지위. 철학은 이 답에 합의하지 않는다. 이 프로젝트에서는 이 논쟁을 지금 해결할 필요가 없다 더 중요한 것은: **도덕심리의 인과적 기원 과 도덕주장의 정당성** 을 분리하는 것. 예: 왜 나는 불공정을 싫어하는가? 와 불공정은 왜 잘못된가? 는 다른 질문이다. 첫 번째는: **CAUSAL EXPLANATION** 두 번째는: **NORMATIVE JUSTIFICATION** 이다. 오히려 윤리는 자연적 성향에 'NO'를 말할 수 있다 나는 복수하고 싶다. 하지만: 하지 않는다. 나는 내집단을 편애하고 싶다. 하지만: 공정하게 판단한다. 즉 윤리의 중요한 특징 중 하나는: **BIOLOGICAL IMPULSE CAN BE NORMATIVELY OVERRIDDEN** 이라는 점이다. 앞 장의 **VALUE STACK**이 다시 양방향으로 작동한다. 이것이 윤리의 핵심 후보다 이라고 단정할 수는 없지만,

매우 강한 후보표현은 가능하다. 윤리는 여러 행위자의 가치가 충돌할 때, 단순히 누구의 욕망이 더 강한지를 묻는 것이 아니라 어떤 이유와 규칙이 서로에게 정당화될 수 있는지를 묻는 층이다. 이 정의는:

- 공감
- 결과
- 권리
- 공정성

을 모두 넣을 공간을 남긴다. 이것은 **문명 학습론**에서도 중요하다 문명이 학습한다. 그런데: 더 효율적으로 착취하는 법 을 배웠다. 그것도 학습이다. 하지만: **moral progress** 라고 부를 수 있는가? 아니다. 따라서: **LEARNING ≠ MORAL PROGRESS** 이다. **MASTER MODEL**에 드디어 새로운 층을 추가한다 기존 **VALUE STACK**: V0 생존 V1 강화 V2 동기 V3 쾌락 V4 느껴지는 가치 V5 Reflection V6 Normativity 이제 V6를 더 엄격히 정의한다. **V6 — INTERPERSONAL NORMATIVITY** 질문: **나와 다른 행위자들의 가치가 충돌할 때, 어떤 이유·규칙·권리배분을 서로에게 요구할 수 있는가?** 이것이 현재 윤리층의 가장 좋은 작동정의다. 그리고 V6에는 최소 네 가지 입력이 필요하다 **OTHERS** 타인의 경험과 이익. **IMPARTIALITY** 내 위치만 특별취급하지 않을 이유. **RIGHTS / CONSTRAINTS** 합계가 좋아도 넘지 말아야 할 경계. **JUSTIFICATION** 그 규칙을 왜 받아들여야 하는가. 따라서: **morality ≠ empathy + cooperation** 이다. 더 정확히: **moral psychology + reflective**

justification 이다. 그렇다면 AI는 윤리적 Reflection을 할 수 없는가 그렇게 단정하지 않는다. 앞장의 기능적 기준을 그대로 적용한다. AI가:

- 여러 행위자의 이해를 표현하고
- 충돌하는 원칙을 비교하고
- 일반화가능성을 검토하고
- 이유를 비판하고
- 기존 규칙을 수정한다면

FUNCTIONAL MORAL REASONING 이라고 부를 수 있을 가능성이 있다. 하지만: 이것이 도덕적 진실을 보장 하지는 않는다. 인간도 마찬가지다. 따라서 VALUE의 전체 구조를 다시 보면 BIOLOGY “나에게 무엇이 중요한가?” ↓ EXPERIENCE “무엇이 좋고 나쁘게 느껴지는가?” ↓ REFLECTION “내가 그것을 추구해야 하는가?” ↓ ETHICS “타인의 가치까지 포함하면 무엇을 서로에게 요구할 수 있는가?” 이 네 층은 상당히 강하다. 매우 중요한 후보명제 윤리의 출발점은 세상에 하나의 목적함수만 존재하지 않는다는 사실일 수 있다. 나에게: 좋다. 당신에게: 나쁘다. 제3자에게: 다른 결과. 따라서: 단일 최적화 로는 해결되지 않는다. 필요한 것은: 가치 간 정당한 관계를 결정하는 규칙 이다. 현재 프로젝트에서 여기까지 잠글 수 있다 ESTABLISHED / STRONG

- 공감·친사회성·공정성·상호성은 인간의 도덕행동과 관련된다.
- 이 능력들은 생물학적·사회적 기원을 가진다.
- 공감은 편향되고 상황의존적일 수 있다.
- 친사회 행동과 도덕적 정당화는 동일하지 않다.

PHILOSOPHICALLY CONTESTED

- 객관적 도덕사실이 존재하는가.
- 어떤 윤리이론이 최종적으로 옳은가.
- 도덕적 이유가 궁극적으로 무엇에 근거하는가.

PROJECT REFRAME 윤리는 개인의 보상함수 확대가 아니라 다중 가치중심 사이의 정당화 문제다. MASTER MODEL v1.4 FROM VALUE TO ETHICS LAYER 1 — SELF SIGNIFICANCE 무엇이 나에게 중요하다. LAYER 2 — FELT VALUE 좋고 나쁘게 느껴진다. LAYER 3 — REFLECTIVE VALUE 나는 어떤 욕망을 승인하는가. LAYER 4 — OTHER MODEL 다른 사람도 경험·목표·이익을 가진다는 것을 표상한다. LAYER 5 — MORAL REASONING 서로 충돌하는 가치에 어떤 규칙을 적용할 것인가. LAYER 6 — NORMATIVE JUSTIFICATION 왜 그 규칙을 누구에게나 요구할 수 있는가. 마지막 층은 아직 철학적으로 열려 있다. HARD LOCKS v1.4 친사회성 ≠ 윤리 공감 ≠ 윤리 협력 ≠ 선험 상호성 ≠ 보편적 의무 공정성 ≠ 하나의 공식 다수결 ≠ 도덕적 정당성 사회규범 ≠ 도덕적 진리 진화적 기원 ≠ 도덕적 정당화 좋게 느껴짐 ≠ 옳음 총효용 증가 ≠ 항상 정당함 학습 ≠ 도덕적 진보 이제 놀라운 질문이 하나 남는다 바둑 AI는: 어떤 수가 승률을 최대화하는가? 를 묻는다. 개인은: 무엇이 나에게 좋은가? 를 묻는다. Reflection은: 내가 무엇을 좋은 것으로 보아야 하는가? 를 묻는다. 윤리는: “여러 존재가 서로 다른 좋은 것을 가질 때, 어떤 규칙 아래 함께 살아야 하는가?” 를 묻는다. 그러면 그 다음 질문은 거의 피할 수 없다. 집단은 하나의 '의사결정자'가 될 수 있는가? 여러 사람의

가치는 하나의 '사회 승률'로 합쳐질 수 있는가 콩도르세 순환 · 애로의 불가능성 · 전략적 투표 · 권리 · 사회후생 개인 최적화에서 집단선택으로 넘어가는 수학적 경계

Arrow의 정리는 바로 이 문제를 엄밀하게 만든 결과다. 충분히 일반적인 개인선호 영역, 사회적 순위의 일관성, 만장일치 존중, 비독재성, 무관한 대안으로부터의 독립성 같은 조건을 모두 요구하면 일반적인 집계규칙을 만들 수 없다. **바둑에서는 왜 이런 문제가 없었는가** 흑과 백이 둔다. 목표는 명확하다. 승리 현재 수 A: 승률 61%. 수 B: 57%. 그러면: $A > B$ 라고 정렬할 수 있다. 물론 평가모델은 틀릴 수 있지만 최소한: **무엇을 기준으로 비교하는가** 에는 합의가 있다. **A와 B** 개인 1: $A > B$ 개인 2: $B > A$ 개인 3: $A > B$ 따라서: $A > B$ 그러면 당연히 $A > C$ 여야 할 것 같다 개별 합리성에서는: $A > B$ 그리고 $B > C$ 라면 $A > C$ 다. 그런데 투표해보면: 개인 1: $A > C$ 개인 2: $C > A$ 개인 3: $C > A$ 따라서: $C > A$ 이다. 결과: $A > B > C > A$ 순환한다. 이것이 콩도르세식 다수결 순환의 핵심이다. 개인들의 순위가 각각 일관적이어도 다수결의 사회적 순위는 비추이적일 수 있다. 먼저 '좋은 규칙'에 무엇을 요구할지 정해야 한다 Arrow의 정확한 정식화는 기술적이지만, 프로젝트에 필요한 핵심 조건은 다음과 같다. **조건 1 UNRESTRICTED DOMAIN** 다양한 개인선호를 허용하자 사회가 미리: 사람들은 반드시 이런 식으로 선호해야 한다. 고 제한하지 않는다. 각 개인이 일관된 순위를 갖는 한 폭넓은 선호구조를 처리할 수 있어야 한다. **조건 2 PARETO / UNANIMITY** 모두가 A를 B보다 원하면 사회도 A를 B보다 원해야 한다 전원이: $A > B$ 인데 사회가: $B > A$ 라고 한다면 이상하다. 따라서 만장일치는 존중하자. **조건 3 NON-DICTATORSHIP** 한 사람의 취향이 사회 전체를 자동 결정하지 않게 하자 어떤 개인 D가: $A > B$ 라고 생각하기만 하면, 다른 모든 사람이 반대해도 항상: $A > B$

가 사회순위가 되어서는 안 된다. **조건 4 INDEPENDENCE OF IRRELEVANT ALTERNATIVES** A와 B의 비교는 기본적으로 A와 B에 대한 개인들의 선호로 결정하자 A와 B 중 누가 더 사회적으로 선호되는지가, 전혀 다른 C가 후보에 들어오거나 빠졌다는 이유만으로 뒤집히지 않게 하자는 직관이다. **조건 5 SOCIAL ORDERING** 사회의 결과도 일관된 순위를 이루자 사회가: $A > B$ $B > C$ $C > A$ 라고 순환해서는 곤란하다. 즉 집단도 최소한: 일관된 선택처럼 행동하길 원한다. 단순히 더하면 안 될까 가장 쉬운 제안은 각 개인의 효용을 합산하는 것이다. 그러면 사회 전체 효용을 최대화하면 된다. 하지만 바로 문제가 생긴다. **이것이 매우 중요한 발견이다** Arrow의 불가능성은: 모든 종류의 사회후생이 절대 불가능하다 는 결과가 아니다. 그 정리는 특정 정보구조와 공리 아래에서 성립한다. 더 많은 정보를 허용하면 다른 사회선택 규칙을 만들 수 있다. 예:

- 효용강도
- 복지수준
- 빈곤
- 건강
- 역량

등. Amartya Sen의 작업이 중요한 이유 중 하나도 사회선택에 사용할 수 있는 정보의 범위를 넓혀 Arrow 이후의 지나친 비판론을 완화한 데 있다. Nobel 위원회 역시 Sen이 개인복지에 관한 정보와 권리 문제 등을 도입해 사회선택이론의 가능성을 크게 확장했다고 평가했다. **사례 병원을 어디에 세울까** 세 후보 지역: A — 도심 B — 외곽 C — 농촌 인구가 많은 지역 사람들은: A를 선호할 수 있다. 다수결만 보면 A. 하지만 농촌 주민은 병원이 없으면: **응급의료 접근 자체가 매우 어렵다**. 그러면 단순 선호순위 외에: **NEED** 정보가 중요해진다. 그래서 사회선택은 단순 투표 문제가 아니다 더 정확하게는: **INFORMA-**

TION + AGGREGATION + RIGHTS + PROCEDURE 문제다. 어떤 정보를 집계할 것인가? 어떤 방식으로 집계할 것인가? 집계해도 침범할 수 없는 영역은 무엇인가? 누가 결정과정에 참여하는가? **Amartya Sen의 자유주의 역설** Sen은 매우 작은 수준의 개인적 권리조차 Pareto 원리와 충돌할 수 있음을 보였다. 이를 너무 거칠게 요약하면: **개인에게 최소한의 자기결정권을 주자 와 모두가 더 선호하는 결과를 언제나 사회가 선택하자** 가 특정 선호구조에서는 동시에 유지되지 않을 수 있다. Nobel 위원회도 Sen의 이 결과가 집단선택과 개인적 권리 사이의 근본적 딜레마를 드러냈다고 설명한다. **이것이 앞 장의 권리 논의를 수학적으로 강화한다** 권리를 단순히: 사회효용을 계산할 때 마이너스 점수를 크게 준다. 로 처리하면 모든 권리가 결국: **교환 가능한 값** 이 된다. 하지만 권리론은 때로 말한다. **이 영역은 집계하지 않는다.** 예: 사적인 신앙. 신체의 일부 선택. 표현. 즉 권리는: **objective의** 향 이라기보다 **domain constraint** 처럼 작동할 수 있다. **이제 더 나쁜 소식** 사람들은 투표규칙을 전략적으로 이용할 수도 있다 지금까지는 사람들이: 솔직하게 자기 선호를 제출한다.

고 가정했다. 하지만 현실에서는 결과를 보고 전략적으로 투표할 수 있다. 예: 내가 가장 좋아하는 후보 A는 당선 가능성이 거의 없다. 그래서: 싫지만 B를 찍어 C를 막자. 가 가능하다. **여기서 AI와 중요한 연결이 생긴다** AI에게 말한다. 사람들의 선호를 파악해서 모두에게 최선인 정책을 만들어라. AI가 아무리 똑똑해도 첫 문제: **실제 선호가 무엇인가?** 사람들이 전략적으로 표현할 수 있다. 둘째: **그 선호들을 어떻게 결합할 것인가?** 집계규칙 선택. 셋째: **선호 자체가 정당한가?** 누군가 차별을 선호할 수도 있다. 넷째: **권리는?** 집계에서 제외해야 할 영역이 있을 수 있다. 따라서 AI의 계산력만으로 문제가 사라지지 않는다. **매우 중요한 HARD LOCK MORE COMPUTE ≠ SOCIAL CHOICE SOLUTION** 이다. 사회선택의 일부 어려움은: **계산량 부족** 이 아니다. **서로 만족하기 어려운 규범적 조건** 때문이다. 컴퓨터가 무한히 빨라도 Arrow 정리의 논리적 충돌은 없어지지 않는다. **예를 들어 의제설정이 결과를 바꿀 수 있다** 앞의 순환: $A > B, B > C, C > A$ 를 생각해보자. 먼저: A vs B 투표. A 승리. 다음: A vs C 투표. C 승리. 최종: C. 그런데: B vs C 먼저 하면? B 승리. 그다음 B vs A. A 승리. 최종: A. 즉: **같은 사람 같은 선호 인데 투표순서 에 따라 결과가 달라질 수 있다.** **이건 AI 추천에서도 똑같다** AI가 사용자에게: A B C 중 고르게 한다. 표면적으로: **사용자가 선택했다.** 그런데 AI가:

- 어떤 후보를 보여줬는지
- 어떤 순서로 보여줬는지
- 기본값을 무엇으로 뒀는지

가 선택에 영향을 준다. 따라서: **CHOICE ≠ FULL AGENCY** 이다. 사회선택에서도: **MENU DESIGN** 가 정치적이다. **이것은 REFLECTION과 연결된다** 시민 A: 세금을 낮추고 싶다. 그런데 토론을 통해: 낮은 세금이 의료접근에 미치는 영향을 몰랐다. 고 배운다. 선호가 바뀐다. 더 나아가: 나는 세금을 단순 개인비용으로만 보고 있었구나. 라고 가치기준을 수정한다. **COLLECTIVE REFLECTION** 후보다. 즉 민주주의에는: **선택 전 학습** 이 존재할 수 있다. **그러면 Arrow 정리는 중요하지만 전체 민주주의 이론은 아니다** Arrow의 프레임에서는 개인의 선호를 입력으로 놓는다. 하지만 민주주의에서는: **선호를 형성하는 과정** 자체도 중요하다. 따라서 숙의민주주의 같은 관점에서는: 주어진 선호를 어떻게 더할까? 뿐 아니라 시민들이 서로 이유를 주고받으며 어떤 선호를 형성할까? 를 묻는다. Arrow 정리의 중요성을 유지하면서도 적용범위를 과장해서는 안 된다. SEP 역시 제한된 선호영역, 추가 정보, 점수·등급, 기수적 정보 등을 사용할 경우 여러 가능성 정리가 열린다고 설명한다. **이 차이는 엄청나게 중요하다** 민주주의에서: 국민의 뜻을 발견한다. 라는 표현은 강하지만 때로

위험하다. 더 엄밀하게는: 특정한 제도 아래에서 시민들의 선호·판단을 공동결정으로 변환한다. 이다. 즉: AGGREGATED OUTCOME $\neq$ PRE-EXISTING GENERAL WILL 이다. 여기서 지금까지의 최적화 연구가 다시 돌아온다 앞 장: 한 목표 아래 최적화. 앞 장: 여러 목표 사이 Pareto frontier. 앞 장: 여러 사람의 가치 충돌. 앞 장: 여러 사람의 선호를 합치는 규칙 자체가 여러 조건 사이 trade-off를 갖는다. 즉 복잡성이 계속 한 층씩 올라갔다. 정확한 전달구조 AI prediction $\downarrow$ consequence map $\downarrow$ individual / group values $\downarrow$

aggregation procedure $\downarrow$ rights / constraints $\downarrow$ legitimate authority $\downarrow$ collective decision 이다. AI가 주로 강해지는 곳은: consequence map 이다. 그 뒤의 단계는 사라지지 않는다. 이것은 바둑의 최적수와 완전히 다르다 바둑에서는: 승리라는 기준에서 최적수. 사회에서는: “어떤 사회선택 규칙을 정당하다고 받아들였을 때의 최적 정책” 일 뿐이다. 따라서: SOCIAL OPTIMUM IS RULE-DEPENDENT 이다. 그래서 헌법은 사회선택의 메타구조로 볼 수 있다 일반정책: 무엇을 결정할까? 헌법: 누가 어떤 절차와 어떤 한계 안에서 결정할까? 이다. 즉: FIRST-ORDER CHOICE 위에 SECOND-ORDER CHOICE ARCHITECTURE 가 있다. 시장도 사회선택 메커니즘 아닌가 어느 정도는 그렇다. 소비자가: 돈으로 투표 한다는 표현이 있다. 하지만 엄밀히는 민주주의와 다르다. 시장에서는 영향력이: 구매력 에 따라 달라진다. 한 사람 한 표가 아니다. 또 공공재·외부효과·권리 문제가 있다. 따라서: MARKET AGGREGATION $\neq$ DEMOCRATIC AGGREGATION 이다. 여기서 문명 학습모델이 한 단계 수정된다 앞 장에서는 사회를: 다중센서 오류수정망 이라고 했다. 이제 추가해야 한다. 사회는 동시에: MULTIPLE AGGREGATORS 를 가진다. 즉:

- 시장은 한 방식으로
- 선거는 다른 방식으로
- 법원은 또 다른 방식으로

분산된 정보를 공동결정으로 바꾼다. 따라서 문명은 '하나의 optimizer'가 아니라는 결론이 더 강해진다 정확히는: 문명은 서로 다른 정보를 감지하고 서로 다른 기준으로 집계하는 여러 의사결정장치가 중첩되어, 서로의 결정을 제한·교정·재협상하는 다중루프 체계다. 이것은 앞 장의: SOCIETY $\neq$ ONE OPTIMIZER 를 수학적으로 보강한다. 그렇다면 집단 합리성은 불가능한가 아니다. 개인의 합리성조차 완벽하지 않다. 사회는:

- 절차를 명시하고
- 서로 다른 제도를 결합하고
- 결과를 수정하고
- 권리를 보호

하면서 작동할 수 있다. 즉: PERFECT SOCIAL RATIONALITY 보다 CORRIGIBLE COLLECTIVE DECISION 이 더 현실적인 목표일 수 있다. 이것이 민주주의의 더 강한 정당화 후보가 된다 민주주의가 항상 최선의 정책을 선택해서가 아니다. 오히려:

- 경쟁정당
- 선거
- 표현의 자유
- 야당

- 법원
- 언론

을 통해: 집단선택 절차 자체를 계속 수정할 가능성 이 있다는 점이 중요할 수 있다. 다시: ERROR CORRECTION 이다. AI 시대에는 새로운 위험이 생긴다 AI가:

- 의제 설정
- 후보 추천
- 정보 요약
- 정책평가
- 여론 예측

까지 담당한다. 그러면 AI는 최종표를 던지지 않아도: PREFERENCE FORMATION + AGENDA FORMATION 에 개입한다. 이는 집단선택 입력 자체를 바꾼다. 앞 장의 agency 모델이 사회로 확장된다 개인 Agency:

- option generation
- evaluation
- choice
- revision.

집단 Agency:

- agenda creation
- public information
- deliberation
- aggregation
- implementation
- revision.

즉 집단 자유 역시: 투표버튼 하나 로 끝나지 않는다. 이 모델이 Arrow보다 큰 이유 Arrow는 특히: 3 → 4 즉 선호를 사회순위로 집계하는 핵심 문제를 엄밀하게 보여준다. 우리 문명 모델은: 집계 전후의 인터페이스까지 포함한다. 따라서: ARROW IS A CENTRAL CONSTRAINT, NOT THE WHOLE MODEL 이다. MASTER HARD LOCKS v1.5 개인 합리성 ≠ 집단 합리성 다수결 ≠ 항상 일관된 사회순위 Arrow 정리 ≠ 민주주의 불가능 불가능성 ≠ 모든 결정규칙이 동등하게 나뉘 선호집계 ≠ 사회적 가치 발견 국민의 뜻 ≠ 집계 이전에 존재하는 단일 선호라고 자동 가정 선호의 합 ≠ 정당성 효용 합계 ≠ 권리 더 많은 계산능력 ≠ 집단선택 문제 해결 예측 정확도 ≠ 정책 정당성 투표 ≠ 선호형성 과정 전체 시장 선택 ≠ 민주적 선택 AI 추천 ≠ 중립적 선택환경 이 장의 가장 중요한 결론 사람이 여럿이 되는 순간 최적화 문제는 단순히 변수의 수가 늘어나는 것이 아니다. '누구의 가치'를 '어떤 규칙'으로 결합할 것인지가 문제 자체의 일부가 된다. 그래서 사회에는: 하나의 자연적으로 주어진 승률표가 없다. 있을 수 있는 것은: 어떤 규범·정보·절차를 채택했을 때의 사회적 평가다. 이것이 사회의 알고리즘과 바둑 알고리즘의 가장 깊은 분기점 가운데 하나다 바둑에서는: 알고리즘이 게임 안에서 움직인다. 사회에서는: 알고리즘 자체가 게임의 규칙이 될 수 있다. 선거제도. 세금제도. 시장규칙.

법절차. AI 추천시스템. 모두: CHOICE ARCHITECTURE 다. 따라서 권력의 아주 깊은 정의가 나온다. 권력은 선택결과를 얻는 능력뿐 아니라, 어떤 선택지가 보이고 어떤 정보가 사용되며 어떤 집계규칙으로 공동결정이 만들어질지를 설계하는 능력이기도 하다. 그리고 다음 질문이 열린다

사회에는 하나의 완벽한 목적함수가 없다. 완벽한 집계규칙도 없다. 그런데 실제 문명은 계속 결정을 한다. 그렇다면: 불가능성과 가치충돌을 안고도 어떻게 '괜찮은 결정'을 만들 수 있는가? 여기서 새로운 개념이 필요하다. LEGITIMACY 결과가 모두에게 최선이 아니더라도: 왜 패배한 사람도 이 결정을 일정 정도 받아들여야 하는가? 이 질문이다.

8장. 정당성과 자기제한

최선의 결과에 합의할 수 없을 때 무엇이 결정을 정당하게 만드는가?

정당성은 결과의 효율성만이 아니라 결정 절차, 기본권, 이유 제시, 권력의 제한, 이의제기와 수정 가능성에서 생긴다. 최선의 결과가 없다면 무엇이 결정을 정당하게 만드는가 절차적 정당성 · 동의 · 이유제시 · 권리 · 속의 · 패배자의 지위 정치철학에서도 정당성은 단순한 지지율이나 실제 동의와 구별된다. 정치권력이 강제를 행사할 권리가 무엇에서 나오는지, 그리고 시민에게 복종의무까지 발생하나는 서로 다른 문제로 논쟁된다. 현대 이론은 절차, 공적 정당화, 정치참여, 결과의 질 등을 서로 다르게 결합한다. 다수결에서 졌다 세금을 올릴지 투표했다. 51%: 올리자. 49%: 반대. 결과: 세금 인상. 여기서 이상한 질문이 생긴다. 49%에게 말한다. 왜 따라야 합니까? 답: 네가 졌으니까. 충분할까? 사회에서는 규칙 자체가 강제된다 나는: 이 세금을 원하지 않는다. 그래도 내야 한다. 나는: 이 법에 반대한다. 그래도 적용된다. 즉 정치적 결정은: 다른 사람의 선택을 실제로 제한 한다. 그래서 필요한 것은 단순: **decision** 이 아니라 **justified coercion** 이다. 정치적 정당성 이론에서 바로 이 점, 즉 정치권력이 법과 정책을 강제할 권리를 어떤 조건 아래 갖는지가 중심 문제다. 따라서 세 개를 분리한다 **CORRECTNESS** 결정이 사실적으로 옳은가. **JUSTICE** 결정 내용이 정의로운가. **LEGITIMACY** 그 결정이 공동체에 권위 있게 적용될 자격이 있는가. 셋은 겹칠 수 있지만 동일하지 않다. SEP의 정당성 논의에서도 절차적으로 정당한 결정이 완전히 정의로운 결정과 동일할 필요는 없다는 구분이 명시적으로 등장한다. 만장일치도 해결책이 아니다 100명 모두의 동의가 있어야 한다고 해보자. 한 사람이 반대하면: 아무것도 결정하지 못한다. 그 한 사람에게 사실상: **veto power** 가 생긴다. 따라서: **CONSENSUS** ↑ → **legitimacy intuition** ↑ 일 수 있지만 **decision capacity** ↓

가 될 수 있다. 다시 trade-off다. 사례 70%가 30%의 재산을 몰수하자고 결정했다 공정한 선거. 한 사람 한 표. 70% 찬성. 이것만으로: 정당하다고 말할 수 있는가? 아마 많은 이론이: 아니다. 라고 할 것이다. 왜냐하면: 무엇을 다수결에 부칠 수 있는가 자체에 경계가 필요하기 때문이다. 여기서 권리가 들어온다 권리는 일부 영역을: **MAJORITY AGGREGATION** 에서 제거한다. 예:

- 신체
- 종교
- 표현
- 적법절차

등. 즉: **RULE OF MAJORITY** 보다 먼저: **RULES OF THE GAME** 가 필요하다. 세 번째 후보 공정한 절차 이제 말할 수 있다. 결과에는 합의하지 못해도, 결정방법에는 합의할 수 있지 않을까? 예:

- 한 사람 한 표
- 정기 선거
- 비밀투표
- 야당 허용

이것이: **PROCEDURAL LEGITIMACY** 다. 순수 절차주의 계열에서는 적절하게 제약된 민주적 절차를 거친 사실 자체가 결정의 정당성을 설명하는 핵심이 된다. 그런데 절차만으로 충분한가 동전을 던진다고 하자. 누구에게도 편파적이지 않다. 50:50. 절차적 공정성만 보면 상당히 훌륭하다. 그런데: 국가 예산 1,000조를 동전으로 결정한다. 고 하면 이상하다. 즉: **FAIR PROCEDURE ≠ COMPETENT PROCEDURE** 이다. 실제로 민주적 정당성 논쟁에서도 단순 공정성만으로 충분하면 추첨 같은 절차와 민주적 숙의를 구별하기 어렵다는 비판이 제기된다. 그러나 전문가에게 모두 맡기면? 가장 뛰어난 전문가 10명이 있다. AI도 있다. 정책예측 정확도 99%. 그들에게 모든 결정을 맡긴다. **EPISTOCRACY** 후보다. 그런데 시민은 묻는다. 그런데 왜 당신들이 나를 지배할 권리가 있습니까? 여기서: **EXPERTISE ≠ AUTHORITY** 가 나온다. AI도 정확히 같다 AI가: 이 정책이 평균소득을 최대화합니다. 라고 완벽히 예측한다고 하자. 그래도: 평균소득을 최우선으로 삼아야 하는가? 누가 희생되는가? 시민이 이 기준을 승인했는가? 가 남는다. 즉: **PREDICTIVE SUPERIORITY ≠ POLITICAL LEGITIMACY** 이다. 여기서 민주주의의 의미가 달라진다 민주주의를: 사람들의 선호를 가장 정확히 더하는 기계 라고만 보면 Arrow 문제에 갇힌다. 다른 정당화: 시민을 정치적 동등자로 대하는 절차 다. 즉: 결과가 내가 원하는 것은 아니지만, 나도 다른 시민과 마찬가지로 그 결정에 참여할 지위를 가졌다. 가 된다. 중요한 단어 **EQUAL STATUS** 한 사람 한 표가 중요한 이유는: 표 하나가 가장 정확한 정보단위라서 만이 아니다. 나와 당신이 정치적으로 동등하다는 상징·제도화 이기도 하다. 따라서 민주주의에는: **epistemic function** 뿐 아니라 **status function** 이 있다.

공적 이유의 핵심 정치적 강제는: 나만 이해하는 사적 이유 가 아니라 함께 살아가는 다른 시민에게도 제시할 수 있는 이유 를 필요로 한다는 생각이다. 공적 이유 이론은 바로 깊은 가치·종교·도덕적 불일치가 존재하는 상황에서도, 공동규칙이 그 규칙의 적용을 받는 자유롭고 평등한 시민들에게 정당화 가능해야 한다는 문제에서 출발한다. 그렇다고 모두 같은 이유에 동의해야 하는가 반드시 그렇지 않다. 사람 A: 환경 때문에 탄소세 찬성. 사람 B: 에너지 안보 때문에 찬성. 사람 C: 산업혁신 때문에 찬성. 같은 결정에: **DIFFERENT REASONS** 로 도달할 수 있다. 따라서: **AGREEMENT ON DECISION ≠ AGREEMENT ON WORLDVIEW** 이다. 공적 이유론 내부에서도 모두가 같은 이유에 합의해야 하는지, 서로 다른 이유가 하나의 결정에 수렴해도 충분한지 논쟁이 있다. 일곱 번째 후보 숙의 투표하기 전에:

- 이유를 말하고
- 반론을 듣고
- 정보를 수정하고
- 선호를 바꾼다.

즉: **DELIBERATION** 이다. 숙의민주주의 계열에서는 시민들이 서로를 동등한 정치공동체 구성원으로 인정하면서 공적 이유를 주고받는 과정 자체가 정당성의 핵심으로 다뤄진다. 사례 쓰레기 소각장 주민: 우리 동네 절대 안 돼. 매우 자연스럽다. 그런데 숙의과정에서:

- 폐기물 처리 필요
- 다른 지역 부담
- 건강위험
- 보상
- 대체입지

를 본다. 결과적으로 반대를 유지할 수도 있다. 중요한 것은: 반드시 마음을 바꾸는 것 이 아니다. 이유구조가 풍부해진다. 그런데 숙의도 이상화하면 안 된다 현실의 토론에는:

- 언변 차이
- 정보격차
- 권력
- 시간
- 사회적 지위

가 작용한다. 말을 잘하는 사람이: 좋은 이유 를 가진 사람보다 이길 수도 있다. 따라서: **DISCUSSION** ≠ **EQUAL DELIBERATION** 이다. 다시 절차 조건이 필요하다 숙의가 정당성에 기여하려면:

- 발언기회
- 정보접근
- 반론 가능성
- 강압 부재
- 이유제시

같은 조건이 중요해진다. 즉: **VOICE** 만이 아니라 **EFFECTIVE VOICE** 다. 여덟 번째 후보 수정 가능성 이제 이 프로젝트 고유의 강한 축이 돌아온다. 아무리 좋은 절차도: 틀릴 수 있다. 그러면 필요한 것은:

- 다음 선거
- 항소
- 재심
- 법 개정
- 헌법심사
- 언론 비판

이다. 즉: **CORRIGIBILITY** 이다. 따라서 패배자의 지위가 핵심이다 민주주의의 질을: 승자가 원하는 것을 얼마나 잘 하는가 로만 보면 안 된다. 더 깊은 시험: 진 사람이 다음 라운드에도 동등한 시민으로 남는가? 이다. 이게 굉장히 중요하다. 그래서 강한 민주주의의 핵심 후보 **LOSERS RETAIN FUTURE OPTIONS** 이다. 앞 장의: **option preservation** 이 정치이론으로 돌아왔다. 좋은 정치제도는 승패를 만들지만: 패배자가 미래의 정치적 선택권 전체를 잃지 않게 한다. 이것이 정당성을 설명하는 강한 이유가 될 수 있다 왜 패자가 결정을 일정 정도 받아들이는가? 단순히: 다수가 이겼으니까 가 아니다. 더 강한 답: 나는 이번 결과에는 반대하지만, 내가 동등한 시민으로 참여했고, 나의 기본권은 유지되며, 결정은 공개적으로 이유가 제시되고, 앞으로 다시 비판·조직·투표·소송을 통해 수정할 수 있기 때문이다. 이 구조가 훨씬 강하다. **LEGITIMACY STACK v1.0** 정당성을 이번 프로젝트에서는 여섯 층으로 잠정 정리할 수 있다. **L1 — INCLUSION** 누가 결정의 영향을 받으며 누가 참여하는가. **L2 — EQUAL STATUS** 참여자들이 동등한 정치적 지위를 갖는가. **L3 — JUSTIFICATION** 결정을 다른 시민에게 이유로 설명할 수 있는가. **L4 — RIGHTS BOUNDARY** 집계가 넘지 못할 기본 경계가 있는가. **L5 — COMPETENCE** 사실·증거·전문성을 처리할 능력이 있는가. **L6 — REVISION** 틀린 결정을 다시 공격하고 수정할 수 있는가. 이것을 사례로 시험한다 감염병 봉쇄정책 정부가 전문가 의견을 받아: 이동을

제한한다. 시민의 자유가 줄어든다. 정당성을 판별하려면? **L2 — Equal Status** 특정 계층만: 예외를 누리고 다른 사람만 강하게 제한된다면 문제가 생긴다. **L4 — Rights** 아무리 공중보전이 중요해도: 제한할 수 있는 범위

가 있다. 예: 법적 근거. 필요성. 비례성. **L6 — Revision** 새로운 증거가 나오면:

- 강화를 하거나
- 완화하거나
- 종료

할 수 있어야 한다. **SUNSET / REVIEW** 구조가 중요하다. 여기서 법치와 민주주의가 정확히 만난다. 민주주의: 누가 결정하는가. 법치: 결정권자도 어떤 규칙과 권리 아래 묶이는가. 숙의: 결정 이유가 무엇인가. 과학·전문성: 사실판단의 질은 어떤가. 이 네 개가 모두 필요할 수 있다. AI는 어디에 들어가는가 AI는 특히: **L5 — COMPETENCE** 를 크게 높일 수 있다.

- 데이터 분석
- 예측
- 시뮬레이션
- 정책 비교

그러나 AI가 L5를 강화한다고: **L1~L4, L6** 이 자동으로 해결되지 않는다. 오히려 AI가 정당성을 해칠 수도 있다 정부: AI가 최적이라고 했습니다. 시민은 모델을 볼 수 없다. 이유도 이해할 수 없다. 수정경로도 없다. 그러면: **EPISTEMIC CAPACITY** ↑ 일 수 있지만 **PUBLIC JUSTIFICATION** ↓ **CHALLENGEABILITY** ↓ 일 수 있다. 즉: **MORE INTELLIGENCE CAN REDUCE LEGITIMACY** 가능성도 있다. 정당성의 중요한 인터페이스 후보 Decision 만 보지 말고: **INPUT RIGHTS** 누가 참여할 수 있는가. ↓ **REASON INTERFACE** 왜 그런 결정인가. ↓ **DECISION AUTHORITY** 누가 결정권을 갖는가. ↓ **RIGHTS CONSTRAINT** 무엇을 결정할 수 없는가. ↓ **APPEAL / REVISION** 어떻게 다시 공격할 수 있는가. 를 봐야 한다. 그래서 정당성과 학습을 하나로 합치면 안 되지만 연결할 수 있다 좋은 학습시스템: 빨리 고친다. 정당한 정치시스템: 누가 어떤 방식으로 고칠 수 있는지도 정당해야 한다. 즉: **EFFECTIVE CORRECTION ≠ LEGITIMATE CORRECTION** 이다. 독재자가 매우 빠르게 정책을 수정할 수도 있다. 하지만 시민을 정치적 동등자로 대하지 않을 수 있다. 이제 패배자의 관점에서 정당성을 정의해보자 정당성 이론은 승자 입장에서 보면 쉽다. 내가 원하는 결과가 나왔다. 문제는 패자다. 정말 강한 테스트: **LOSER'S TEST** 이다. 질문: 내가 이 결정에 끝까지 반대하면서도, 왜 이 결정이 지금 공동체를 대표해 집행될 수 있다는 점은 인정할 수 있는가? 이것이 정당성의 아주 강한 프로젝트 정의를 만든다

정당성은 모두가 결과를 좋아하게 만드는 성질이 아니라, 지속적인 가치충돌 속에서도 패배한 사람을 정치공동체 밖으로 밀어내지 않은 채 공동결정을 만들고 강제하고 다시 수정할 수 있게 하는 제도적 관계다. 이 정의는:

- 다수결
- 권리
- 숙의

- 법치
- 수정 가능성

을 하나로 묶는다. **MASTER MODEL v1.6 FROM OPTIMIZATION TO LEGITIMACY STAGE 1 — PRIVATE OPTIMIZATION** 나는 무엇을 원하는가? **STAGE 2 — SOCIAL CHOICE** 서로 다른 욕망을 어떻게 공동결정으로 바꿀까? **STAGE 3 — ETHICS** 어떤 이유로 서로를 제한할 수 있는가? **STAGE 4 — LEGITIMACY** 불일치가 남아도 누가 어떤 절차로 강제결정을 내릴 수 있는가? **STAGE 5 — CORRIGIBILITY** 그 결정이 틀렸을 때 어떻게 다시 열 수 있는가? 그리고 '좋은 정치'를 새로운 방식으로 정의할 수 있다. 좋은 정치: 모두가 행복한 결정을 찾는 것 만이 아니다. 그것은 불가능할 수 있다. 더 현실적인 후보: 깊은 불일치를 폭력이나 영구배제 없이 잠정적 공동결정으로 변환하고, 패자가 미래의 선택권을 유지하며, 새로운 증거와 이유에 따라 그 결정을 다시 열 수 있게 하는 것. 이것은 최적화보다 훨씬 다른 종류의 문제다. 이것은 매우 큰 정치적 통찰이다. 바둑에서는 상대의 선택권을 줄이는 것이 최고의 전략이다. 정치에서는 상대의 정치적 선택권을 완전히 제거하면: 게임 자체가 무너질 수 있다. 따라서: 전략적 승리 와 제도적 자기보존 이 충돌한다. 그래서 헌법과 규범은 정치게임의 메타계약이다. 정치인은 승리하고 싶다. 하지만:

- 선거결과 존중
- 야당 존재
- 표현 자유
- 권력 이양

같은 규칙이: 가능한 전략 일부를 금지 한다. 즉: **CONSTITUTIONAL CONSTRAINTS PRESERVE THE GAME** 이다. 그래서 정치체계의 가장 위험한 순간 한 세력이 승리한다. 그리고: 다시 질 가능성 자체를 제거하려 한다. 이때: 일회성 정치승리 ↓ 영구적 선택권 비대칭 으로 바뀐다. 앞 장의 Agency, 앞 장의 Choice Architecture, 앞 장의 Legitimacy가 하나로 연결된다. 이 장의 **HARD LOCKS** 정당성 ≠ 인기 정당성 ≠ 실제 동의 정당성 ≠ 다수결 정당성 ≠ 정책 정확성 전문성 ≠ 통치권 공정한 절차 ≠ 충분한 절차 숙의 ≠ 합의 패배 ≠ 정치적 배제 정당성 ≠ 정의 적응적 통치 ≠ 정당한 통치 AI 정확도 ≠ AI의 정치적 권위 설명가능성 ≠ 이의제기 가능성 이 장의 가장 강한 결론 사회에서는 완벽한 승률표도, 완벽한 집계규칙도, 완벽한 정책예측도 최종 답이 아니다. 왜냐하면 공동결정은 결국: 사람에게 강제 되기 때문이다. 그래서 질문은: 어느 정책이 가장 좋은가? 에서 끝나지 않는다. 왜 당신이 반대하는데도 우리가 이 결정을 당신에게 적용할 권리가 있는가? 까지 가야 한다. 그리고 현재 살아남은 가장 강한 답은: 참여 + 동등한 지위 + 공적 이유 + 기본권 + 판단능력 + 수정가능성 의 결합이다. 이번 노드는 헌법을 '좋은 규칙집'이 아니라 권력자가 미래의 자기 자신까지 묶는 장치로 보겠습니다. 핵심은 단기 승리와 장기 게임의 보존 사이의 긴장입니다. 정치가 게임이라면 왜 승자가 게임을 파괴하지 않아야 하는가 헌법 · 자기구속 · 반복게임 · 신뢰 · 권력분립 · 제도의 생존

헌법적 자기구속의 오래된 난점은 바로 이것이다. 정치공동체가 현재의 다수나 지도자에게 불편한 헌법규칙을 왜 계속 지키는지 설명하려면 단순한 '율리시스가 뚝대에 자신을 묶었다'는 비유를 넘어 실제 집행 메커니즘을 설명해야 한다. 먼저 가장 이상한 질문 선거에서 이겼다. 의회 다수. 대통령. 행정부. 국민 지지도 높다. 그렇다면 왜 스스로 말해야 하는가? 여기까지는 할 수 있지만 여기부터는 할 수 없다. 왜? 권력이 있는데. 이게 중요한 이유 우리는 흔히: 민주주의자는 민주주의를 존중하니까 규칙을 지킨다. 고 설명한다. 물론 규범은 중요하다. 하지만 제도이론에서는 더 어려운 질문을 해야 한다. 규칙을 깨면

이득인 사람에게 왜 규칙 준수가 지속되는가? 이다. 이것이: **CREDIBLE COMMITMENT PROBLEM** 이다. 권력분립의 새로운 의미 흔히: 권력을 세 개로 나눈다. 행정부, 입법부, 사법부. 라고 배운다. 하지만 기능적으로 더 중요한 것은: **한 행위자가 모든 오류수정 인터페이스를 동시에 장악하지 못하게 하는 것** 이다. 왜 모든 권력이 한곳에 있으면 위험한가 지도자가:

- 정책을 만든다.
- 집행한다.
- 결과를 평가한다.
- 위법 여부를 판단한다.
- 비판자를 처벌한다.

그러면: **ERROR SIGNAL** 이 지도자를 통과해야 한다. 지도자가 틀리면? 그 오류를 고치는 장치도 같이 틀릴 수 있다. 즉: **SINGLE POINT OF INSTITUTIONAL FAILURE** 가 된다. 앞 장의 **문명학습과 정확히 연결된다** 문명 학습의 4층: 탐색 → 감지 → 판별 → 수정 그런데 한 권력자가:

- 탐색범위
- 감지정보
- 판별방식
- 수정권한

을 모두 통제하면? 학습루프가 닫혀 버릴 수 있다. 따라서: 이라는 강한 후보가 생긴다. **하지만 이것만으로는 부족하다** 헌법에: 대통령 권력을 제한한다. 라고 쓰면 끝인가? 아니다. Daryl Levinson이 지적한 헌법적 약속의 핵심 난점도 바로 이것이다. **종이가 어떻게 권력을 묶는가?** 이다. 헌법은 **밧줄이 아니다** 울리시스는 실제 밧줄에 묶였다. 정치인은 헌법책에 물리적으로 묶이지 않는다. 따라서: **CONSTITUTIONAL RULE** 이 실제 제약이 되려면: **ENFORCEMENT COALITION** 이 필요하다. 즉 누군가가 규칙을 지킬 유인과 능력을 가져야 한다. 따라서 **헌법의 강도** **written constraint** 가 아니라 **enforceable constraint** 로 측정해야 한다. 즉: **RULE + INDEPENDENT CAPACITY TO ENFORCE RULE** 가 필요하다. 권력제약 연구 역시 규칙의 설계만이 아니라 그것을 실제로 집행할 기관이 중요하다고 강조한다. **매우 중요한 원리 CREDIBILITY MAY REQUIRE GIVING AWAY POWER** 이다. 즉: **권력을 일부 잃어야 더 강한 약속을 할 수 있다.** 처음 보면 역설이다. **이건 앞 장의 option value와 흥미롭게 충돌한다** 우리는 계속: **선택권 보존** 을 중요하게 봤다. 그런데 여기서는: **일부 선택권을 스스로 제거 하는 것이 가치 있다.** 왜? 나의 미래 행동에 대해 다른 사람이 예측할 수 있기 때문이다. 즉: **OPTION VALUE** 와 **COMMITMENT VALUE** 가 충돌한다. 따라서 **민주주의의 안정은 반복게임으로 볼 수 있다** 오늘: A 승리. 내일: B가 이길 수도 있다. A는 오늘 B를 완전히 제거할 수 있다. 하지만 그러면 B도 미래에: 내가 기회가 오면 A를 제거해야겠다. 고 생각한다. 결과: **winner-take-all escalation** 이다. 이것이 **반복게임의 핵심 직관**이다 한 번만 만나는 게임:

최대한 가져가라. 반복게임: 오늘의 행동이: **내일 상대의 행동** 을 바꾼다. 따라서 평판과 신뢰가 가치가 된다. **간단한 그림** 한 번의 권력게임: Defect > Cooperate 일 수 있다. 하지만 반복되면: 오늘 규칙 준수 ↓ 상대도 미래 규칙 준수 기대 ↓ 폭력적 경쟁 비용 ↓ ↓ 장기 기대가치 ↑ 가 될 수 있다. 즉: **SHORT-RUN DOMINANCE ≠ REPEATED-GAME OPTIMUM** 이다. 그래서 헌법은 **패자만 보호하는 것이 아니다** 오늘: A가 권력. 헌법은 B를 보호. 내일: B가 권력. 같은 헌법이 A를

보호한다. 따라서: **constitutional constraint** 은 현재 권력자에게 비용이지만: **future loser insurance** 이기도 하다. 이것이 약한 세력이 규칙을 좋아하는 이유를 설명한다 현재 권력이 약하다. 미래가 불확실하다. 그러면: **강한 제도적 제약** 이 매력적일 수 있다. 권력제약 연구에서도 지도자가 초기에 상대적으로 약한 경우 엘리트에게 실질적 권력접근을 허용해 미래 약속을 신뢰 가능하게 만드는 유인이 커질 수 있음이 게임이론적으로 제시된다. 이것이 헌법의 시간문제다 규칙을 만드는 순간에는: **미래에 누가 승자인지 모름**. 그러면 공정한 규칙을 선호할 수 있다. 하지만 결과가 나온 뒤에는: **승자가 규칙을 바꾸고 싶어짐**. 따라서 헌법은: **EX ANTE COMMITMENT AGAINST EX POST TEMPTATION** 이다. **HARD LOCK POLITICAL COMMUNITY** ≠ **SINGLE PERSON** 이다. 개인의 자기통제와 헌법적 자기제약은: **commitment structure** 가 닮았을 뿐 동일하지 않다. 왜냐하면 미래의 시민은: **다른 사람들** 일 수도 있기 때문이다. 그래서 헌법은 자기구속인 동시에 타인구속이다 현재 세대가: 미래 세대도 이 규칙을 따라라. 라고 한다. 그러면 질문: **왜 과거가 미래를 묶을 권리가 있는가?** 가 생긴다. 즉:

CONSTITUTIONAL COMMITMENT 에는 장점뿐 아니라 민주적 문제가 있다. **문명 학습 모델과 정확히 연결된다** 규칙이 너무 쉽게 바뀌면: **신뢰** ↓ 규칙이 전혀 안 바뀌면: **학습** ↓ 따라서 좋은 헌법은: **STABILITY + CORRIGIBILITY** 를 동시에 필요로 한다. 즉 헌법도 최적점이 아니라 견고한 영역을 찾는다 **FLEXIBLE ENOUGH TO LEARN** 하지만 **RIGID ENOUGH TO COMMIT** 이어야 한다. 앞장의: **robustness** 가 제도설계로 돌아왔다. **아주 중요한 구분 POLICY CHANGE** 와 **GAME-RULE CHANGE** 는 다르다. 정권이: 세율을 바꾼다. 정상적 정치경쟁. 하지만: 다음에 질 가능성을 줄이기 위해 선거·사법·언론 규칙을 바꾼다. 이는: **META-GAME INTERVENTION** 이다. 따라서 민주주의 후퇴를 결과만으로 보면 안 된다 질문: 이 정책이 좌파인가 우파인가? 보다 먼저: **이 결정이 미래 경쟁의 공정성을 바꾸는가?** 를 봐야 한다. 즉: **POLICY CONTENT** 와 **COMPETITIVE ARCHITECTURE** 를 분리해야 한다. 이를 하나의 지표 후보로 만들 수 있다 정치결정 X가 있다. 질문: **이 결정 이후 현재 야당이 동일한 규칙 아래 미래에 집권할 현실적 가능성이 유지되는가?** YES: 정상 정치경쟁 가능성 높음. NO: 민주적 자기잠금 위험. 이것은 매우 강한 진단질문이다. 여기서 이세들의 68수가 다시 등장한다 우리가 초반에 봤다. 이세들의 사후 회고를 이 프로젝트의 언어로 해석하면, 좋은 수는 단순히: 지금 얻는 집 이 아닐 수 있다. **상대의 미래 선택공간을 바꾼다**. 정치에서도 가장 강한 권력행동은: **현재 정책을 이기는 것 보다 미래 상대의 선택공간을 줄이는 것** 일 수 있다. 왜냐하면 민주주의의 목적은 **상대 제거가 아니기 때문이다** 최소한 제도적 수준에서:

CONTESTATION MUST CONTINUE 해야 한다. 따라서 민주주의의 메타목표는: **victory under preserved competition** 이다. 즉: **이기되 경쟁 가능성은 남긴다**. 여기서 신뢰가 핵심 상태변수가 된다 야당이 믿는다. 우리가 선거에서 저도 감옥에 가지 않는다. 집권당이 믿는다. 우리가 다음에 저도 기본권은 남는다. 시민이 믿는다. 정권이 바뀌어도 계약·법질서가 완전히 무너지지 않는다. 그러면: **political stakes** ↓ 갈등이 줄어들 수 있다. 반대로 **선순환도 가능하다** 평화로운 패배 ↓ 패자의 권리 보존 ↓ 다음 세력의 신뢰 ↑ ↓ 패배 수용 비용 ↓ ↓ 다음 권력교체 가능 ↓ **TRUST ACCUMULATION** 이다. 이것은 일종의: **institutional capital** 로 볼 수 있다. 따라서 제도는 규칙 + 역사다 정확히: **Institution** Formal Rules + Enforcement Capacity + Repeated Expectations 으로 보는 것이 강하다. 종이에 동일한 헌법이 있어도: **실제 정치균형** 은 다를 수 있다. 이것이 헌법을 단순 **법률텍스트 이상으로 만든다** 헌법: **coordination focal point** 역할을 할 수 있다. 사람들이: 이런 상황에서는 이렇게 행동해야 한다. 고 서로 기대하게 만든다. 최근 제도이론에서도 헌법적 단계나

초점점이 행위자들의 기대를 조정할 수 있지만, 계획이 실제로 유지되려면 참여자들의 유인과 균형이 맞아야 한다는 점이 강조된다. 그래서 진짜 제약에는 권력이 필요하다 역설: 권력을 제한하려면 다른 권력이 필요하다. 법원이 대통령을 제한하려면: 법원에게:

- 독립성
- 임기
- 집행력

이 필요하다. 의회가 행정부를 견제하려면:

- 예산권
- 조사권

같은 실제 권한이 있어야 한다. 즉: CHECK WITHOUT POWER ≠ CHECK 이다. 따라서 권력분립은 권력을 없애는 것이 아니다 POWER AGAINST POWER 다. 한 권력을: 무권력 으로 만드는 것이 아니라 여러 권력을 만들어: 서로를 제한 한다. 또 trade-off다 DECISIVENESS vs CONSTRAINT 위기 때 빠른 결정을 원한다. 그러나 권력집중은 오남용 위험. 평시에는 느린 견제가 안전. 즉: 제도의 좋은 정도는 환경에 따라 달라질 수 있다. 그러므로 헌법설계도 단일 최적해가 아니다 국가마다:

- 역사
- 정당체계
- 사회분열
- 행정역량

이 다르다. 따라서: UNIVERSAL PERFECT CONSTITUTION 은 현재 프로젝트에서 거부한다. 중요한 것은 기능:

- 권력교체
- 오류수정
- 권리보호
- 약속 신뢰성
- 정책능력

을 어떻게 균형 잡는가다. 헌법을 문명 학습론 안에 넣으면 헌법은: 학습 결과 이면서 학습 조건 이다. 과거 실패: 독재. 권력남용. 내전. ↓ 헌법 규칙 형성. 그 헌법이: 다음 정치학습의 경로 를 제한한다. 즉: MEMORY THAT CHANGES FUTURE SEARCH SPACE 다. 그러나 저장된 기억도 틀릴 수 있다 과거 위기에 대한 과잉반응. 너무 강한 금지. 시대가 바뀔. 그러면: 헌법 자체가 오류 일 수도 있다. 따라서 다시: amendment 가 필요하다. 문명 메타학습: 규칙 수정 규칙 이다. 이 모델의 병목도 명확하다 FAILURE A — Constraint Without Enforcement 헌법만 있음. 집행자 없음. ↓ PAPER CONSTITUTION FAILURE B — Enforcement Capture 법원·선거기관·언론이 집권세력에 종속. ↓ FEEDBACK COLLAPSE FAILURE C — Winner Entrenchment 승자가 경쟁규칙 자체를 변경. ↓ FUTURE OPTION DESTRUCTION FAILURE D — Excessive Rigidity 잘못된 규칙을 고칠 수

없음. ↓ LEARNING FAILURE FAILURE E — Excessive Fragmentation 아무도 결정을 못 함.
 ↓ CAPACITY FAILURE 이 장에서 가장 중요한 HARD LOCKS 헌법문 ≠ 실제 제약 약속 ≠ 신뢰
 가능한 약속 자기구속 ≠ 선의 권력분립 ≠ 권력 제거 제약 증가 ≠ 항상 좋은 정부 선거 승리 ≠ 게임규칙
 변경 권한의 무제한 부여 정책 변경 ≠ 경쟁구조 변경 현재 다수 ≠ 영구 다수 선택권 보존 ≠ 모든 선택권
 유지 헌법 안정성 ≠ 수정 불가능성 반복게임 유사성 ≠ 정치가 단순 게임 이 장의 가장 강한 결론

헌법이 하는 가장 깊은 일 가운데 하나는 오늘의 승자가 미래의 모든 상태를 독점하지 못하게 만들어,
 오늘의 패자가 패배를 받아들일 수 있게 하는 것이다. 즉 헌법은 단순히 권력을 줄이는 것이 아니다.
 현재 권력의 선택권 일부를 제한해서 미래 모든 행위자의 선택권을 보존한다. 이것은 앞 장의 option
 preservation이 제도수준에서 완전히 새로운 형태로 돌아온 것이다. 그리고 가장 흥미로운 역설 자유를
 유지하려면 일부 자유를 제한해야 한다. 정치인의: 원하는 대로 권력을 쓸 자유 를 제한한다. 그래야
 시민과 야당의: 미래에 선택하고 경쟁할 자유 가 유지된다. 즉: LOCAL FREEDOM CONSTRAINT
 → SYSTEMIC FREEDOM PRESERVATION 이다. 바둑에서 시작한 질문도 또 뒤집힌다 바둑: 좋은
 수는: 상대의 미래 선택권을 줄인다. 민주주의: 좋은 메타규칙은: 승자의 미래 전략 일부를 줄여
 패자의 미래 선택권을 보존한다. 즉: 제도는 플레이어를 최적화하는 것이 아니라 게임의 지속가능성을
 최적화하는 층 이라고 볼 수 있다. 단, '최적화'는 여기서 분석적 표현이지 하나의 사회목적함수가 있다는
 뜻은 아니다. NEXT NODE 제도는 언제 살아남는가 신뢰 · 경로의존 · 자기강화 · 임계점 · 제도붕괴 ·
 회복력 이제 질문은 규칙 설계에서 더 깊은 동학으로 간다. 같은 헌법을 가진 두 나라가 있다고 하자. 왜
 한 곳에서는: 선거 패배 → 권력 이양 → 신뢰 강화 가 반복되고, 다른 곳에서는: 규칙 위반 → 무처벌 →
 더 큰 위반 → 제도붕괴 가 일어나는가? 즉 다음에는: 제도가 한 번 만들어지는 것보다 왜 어떤 제도는
 자기강화되고 어떤 제도는 자기붕괴하는가 를 봐야 한다. 여기서:

- 반복게임
- 기대
- 경로의존
- 임계점
- 복원력
- 문명 학습

이 다시 하나의 동역학으로 결합된다. 그리고 이 단계에 도달하면 우리는 정치제도를 단순한 '규칙집'이
 아니라: 사람들의 기대를 계속 생산하면서 자기 자신을 유지하거나 파괴하는 동적 시스템 으로 다룰 수
 있게 된다. 이번 노드의 핵심은 제약이 약함의 반대가 아니라 때로 신뢰를 생산하는 자산이라는 점입니다.
 실제 권력공유 연구에서도 약속을 신뢰 가능하게 만들려면 통치자가 다른 행위자에게 실제 국가권력을
 나눠 자신의 향후 배신 능력을 줄여야 할 수 있습니다.

9장. 제도와 사회적 실재

만질 수 없는 규칙과 지위는 어떻게 실제 행동을 바꾸는가?

제도는 머릿속 믿음만도 물리적 사물만도 아니다. 반복되는 기대, 규칙, 기록, 권한, 제재, 물질적 기반이 서로 지지할 때 사회적으로 실재한다. 제도는 언제 살아남고 언제 무너지는가 기대·경로의존·자기강화·임계점·복원력·제도붕괴 제도의 실재는 규칙문보다 반복되는 기대와 행동에 있다. 같은 헌법 두 권 두 나라에 거의 같은 조문이 있다고 하자. 둘 다:

- 선거
- 법원 독립
- 표현의 자유
- 권력분립

을 적어놓았다. 그런데 한 나라에서는: 대통령이 판결을 따른다. 다른 나라에서는: 판결을 무시한다. 첫 나라에서는 야당이 선거에서 져도 다음 선거를 기다린다. 두 번째에서는: 어차피 다음에도 공정하지 않을 것이다. 라고 생각한다. 종이에 적힌 규칙은 비슷하다. 그러나 실제 제도는 전혀 다르다. 따라서: **FORMAL RULE ≠ OPERATING INSTITUTION** 이다. 제도에는 최소 세 층이 있다 ① **RULE** 무엇을 해야 하는가. 예: 선거에서 진 정부는 권력을 넘긴다. ② **EXPECTATION** 다른 사람도 그 규칙을 따를 것이라고 믿는가. 상대도 지면 넘길 것이다. ③ **ENFORCEMENT** 위반하면 실제 비용이 있는가. 법원·의회·시민·정당이 저항할 수 있다. 이 세 가지가 결합되어야 제도가 작동한다. 따라서 제도는 단순한 명령이 아니다 정지신호를 생각하자. 법에는: 멈춰라. 라고 쓰여 있다. 그런데 모두가: 아무도 안 멈춘다. 고 믿으면? 혼자 멈추는 것이 오히려 위험해질 수 있다. 즉 규칙의 효과는: 내가 규칙을 어떻게 보는가 뿐 아니라 다른 사람들이 규칙을 어떻게 볼 것이라고 내가 믿는가 에 달려 있다. 이것이 '제도가 살아 있다'는 뜻이다 살아 있다는 생물학적 의미가 아니다. 분석적으로: 규칙이 사람들의 미래행동 예측에 실제로 들어가 있다. 는 뜻이다. 예: 기업이 계약을 맺을 때: 법원이 계약을 집행할 것이다. 라고 믿는다. 그래서 투자한다. 즉 법은 판결이 난 뒤에만 작동하는 것이 아니다. **판결 이전의 행동을 바꾼다. 이게 앞 장의 커밋먼트와 정확히 연결된다** 약속: 나는 이렇게 하겠다. 신뢰 가능한 제도: 내가 다르게 행동하면 실제 결과가 생긴다. 따라서 제도는: **미래 행동에 관한 기대를 현재 선택에 묶는다.** 가장 단순한 자기강화 루프 규칙 준수 ↓ 신뢰 증가 ↓ 규칙 준수의 기대수익 증가 ↓ 다음 준수 증가 ↓ 더 높은 신뢰 이다. 즉: COMPLIANCE → EXPECTATION → COMPLIANCE

라는 양의 되먹임이다. 이게 제도의 자기강화다 중요: **SELF-REINFORCING ≠ SELF-GUARANTEEING** 이다. 한 번 강해졌다고 영원히 유지된다는 뜻은 아니다. 단지: **현재 작동이 미래 작동 가능성을 높이는 메커니즘** 이 존재한다는 뜻이다. 언어도 비슷하다 어떤 단어가 의미를 갖는다. 왜? 많은 사람이 같은 방식으로 사용하기 때문이다. 혼자: 이 단어는 오늘부터 다른 뜻이다. 라고 해도 사회적으로 작동하지 않는다. 즉 제도·화폐·언어에는 모두: **SHARED EXPECTATION** 이 중요한 면이 있다. 동일하다는 뜻은 아니다. 붕괴 루프 규칙 위반 ↓ 처벌 없음 ↓ 규칙 유효성에 대한 기대 하락 ↓

다른 행위자의 준수 감소 ↓ 위반 증가 ↓ 추가 기대 하락 이다. 즉: **DEFECTION → EXPECTATION COLLAPSE → MORE DEFECTION** 이다. 사례 부패 처음: 공무원 1명이 뇌물을 받는다. 처벌된다. 신호: **규칙 살아 있음**. 다른 경우: 공무원 1명이 받는다. 처벌되지 않는다. 다른 공무원: 안 받는 나만 손해 아닌가? 라고 생각하기 시작한다. 그러면 부패는: **개인의 도덕성 문제** 에서 **균형 문제** 로 바뀔 수 있다. 여기서 **임계점이 등장한다** 처음 몇 번의 위반은 제도를 흔들지 않을 수 있다. 사람들은: 예외적인 사건. 이라고 본다. 하지만 어느 순간: 이제 예외가 아니라 규칙이 바뀐 것 같다. 고 생각할 수 있다. 그러면 행동이 급격히 변한다. 즉: **NONLINEAR CHANGE** 가 가능하다. 이것이 **임계점의 핵심** 직관 위반률이: 10% → 11% 일 때 변화는 작지만, 어떤 기대 임계값을 넘으면: 40 → 80 처럼 행동이 급변할 수 있다. 왜냐하면 사람이 단순히 위반 자체가 아니라: **다른 사람의 미래행동에 대한 추정** 을 바꾸기 때문이다. 여기서 **경로의존이 생긴다** 같은 현재 조건이라도: **어떻게 여기까지 왔는가** 에 따라 결과가 달라질 수 있다.

사회 A: 오랜 권력교체 경험. 사회 B: 반복적 쿠데타 경험. 현재 헌법문이 같아도: **기대가 다르다**. 따라서 현재 상태만으로 미래를 설명하기 어렵다. 이것은 **개인의 습관과 닮았다** 처음 담배 한 개비. 아주 작은 선택. 반복. 습관·의존·환경 변화. 몇 년 후: 그냥 안 피우면 되지. 가 어려워진다. 왜? 과거 행동이: **현재 선택비용** 을 바꿨기 때문이다. 제도도 비슷할 수 있다. 따라서 **앞 장의 커밋먼트 구조가 다시 등장한다** 과거 선택: **future option set** 을 바꾼다. 즉 제도도: Past choices ↓ Current constraints ↓ Future probabilities 라는 시간구조를 가진다. **복원력에는 두 종류가 있다** *TYPE A — BOUNCE BACK* 원래 상태로 돌아간다. 예: 선거 분쟁 후 결과 인정. *TYPE B — ADAPTIVE RECOVERY* 원래 상태가 문제였다면: 규칙을 수정하면서 핵심 기능은 유지한다. 예: 선거법 보완. 감시기관 강화. 즉: **RESILIENCE ≠ STATUS QUO PRESERVATION** 이다. 오히려 너무 경직된 제도는 **복원력이 낮을 수 있다** 충격이 왔다. 규칙: **절대 변경 불가**. 결과: 현실과 제도의 괴리 확대. 그러면 한 번에 깨질 수 있다. 따라서 복원력에는: **controlled adaptability** 가 필요하다. 좋은 제도는 무엇을 보존해야 하는가 모든 세부규칙을 보존할 필요는 없다. 오히려: **CORE FUNCTION** 을 보존해야 한다. 민주주의라면:

- 경쟁
- 권력교체
- 기본권
- 이의제기
- 수정가능성

이 더 중요하다. 선거방식의 세부는 바뀔 수 있다. 즉: **RULE PRESERVATION ≠ FUNCTION PRESERVATION** 이다. 이제 제도의 '질병'을 말할 수 있는가 조심해야 한다. 하지만 분석적으로는: **기능 실패** 를 정의할 수 있다. 예: 선거는 열림. 그러나 경쟁 없음. 법원 존재. 그러나 독립 없음. 언론 존재. 그러나 비판 불가. 즉: **STRUCTURE PRESENT / FUNCTION ABSENT** 이다. 사례 법원 Installed: 법원 건물과 법률 존재. Accessible: 시민이 실제 소송 가능. Effective: 법원이 행정부를 실제로 제약할 수 있음. Output: 권리침해가 실제로 교정됨. 따라서: **COURT EXISTS** 와 **RULE OF LAW WORKS** 는 다르다. 이 모델은 **제도붕괴를 더 일찍 감지하게 한다** 공식 제도 폐지: 아주 늦은 신호 일 수 있다. 먼저 흔들리는 것은:

- 접근성
- 독립성

- 집행력
- 기대

일 수 있다. 즉: **FUNCTIONAL DEGRADATION PRECEDES FORMAL COLLAPSE** 이다. 그래서 **제도 건강을 법조문만으로 측정하면 안 된다** 봐야 할 것은: 규칙 무엇이 쓰여 있는가. 권한 누가 실제로 집행할 수 있는가. 이용가능성 시민이 실제로 접근 가능한가. 기대 사람들이 실제로 믿는가. 결과 규칙이 실제 행동을 바꾸는가. 이다. **신뢰가 항상 좋은 것은 아니다** 매우 중요하다. 부패한 조직을: 믿는다. 고 해서 좋은 사회가 아니다. 따라서: **TRUST ≠ INSTITUTIONAL QUALITY** 이다. 더 중요한 것은: **WARRANTED TRUST** 즉 실제 성과와 집행구조에 근거한 신뢰다. 이것은 아주 중요한 **문명적 구조**다 사람이 선택해서: **신뢰** 하는 것이 아니다. 사람이 틀릴 수 있다고 가정하고: **검증장치** 를 만든다. 그 검증장치가 잘 작동하는 것이 반복되면: **더 높은 신뢰** 가 생긴다. 즉: **distrust assumption** → checks → predictable behavior → warranted trust 이다. **법도 같다** 판사: 나는 공정한 사람입니다. 보다:

- 이유제시
- 공개재판
- 항소

가 더 강하다. 즉: **PERSONAL VIRTUE** 보다 **CHECKABLE PROCEDURE** 가 제도신뢰를 만든다. 이것은 **AI 거버넌스**에도 똑같다 AI가: 저는 안전합니다. 라고 말하는 것은 의미가 약하다. 필요한 것은:

- 독립 평가
- 감사
- 로그
- 이의제기
- 롤백

이다. 즉: **TRUST THE SYSTEM** 보다 **MAKE THE SYSTEM AUDITABLE** 이 중요하다. 따라서 **강한 문명은 최대효율 시스템이 아닐 수 있다** 후보: 약간의 비효율 + 대체경로 + 예비역량을 가진 시스템. 예:

- 비상재고
- 여러 에너지 공급원
- 권력분립
- 독립기관

즉: **SLACK CAN HAVE OPTION VALUE** 이다. **제도에도 '여유'가 있다** 예: 독립언론 여러 곳. 정당 여러 개. 여러 법적 구제수단. 평소에는: 중복. 처럼 보일 수 있다. 하지만 한 경로가 막히면: **다른 경로가 오류신호를 운반** 할 수 있다. 이제 **제도붕괴의 깊은 정의**가 나온다 제도가 무너지는 것은: 규칙을 한 번 어기는 것 이 아니다. 더 강한 정의: 위반 이후에도 그 위반을 감지·전달·제재·수정할 대체 경로들이 연쇄적으로 사라지고, 행위자들이 더 이상 규칙 준수를 합리적 기대라고 보지 않게 되는 상태 이다. 즉, 붕괴는 피드백 구조의 실패와 기대의 전환이 결합한 상태다. 이것은 갑작스러운 붕괴를 설명한다 겉보기에는 멀쩡하다. 그러다 어느 사건에서:

- 법원 무시
- 선거불복
- 군 이탈
- 금융패닉

등이 발생. 사람들의 기대가 동시에 바뀐다. 그러면: **빠른 체제전환** 이 가능하다. 즉 물리적 시설이 하루 만에 사라진 것이 아니다. **EXPECTATION REGIME** 이 바뀐 것이다. 따라서 사회에서는 **예측이 단순 관찰이 아니다** 사람들이 믿는 미래가: **현재 행동을 바꾸고 그 미래를 실제로 만들 수 있다.** 즉: **REFLEXIVITY** 가 있다. 이걸 금융시장과 정치에서 특히 중요하다. **MASTER MODEL v1.8 INSTITUTIONAL SURVIVAL LOOP 1. FORMAL RULE** 규칙이 존재한다. ↓ **2. ENFORCEMENT CAPACITY** 위반을 실제로 제한할 수 있다. ↓ **3. EXPECTATION** 다른 사람도 규칙을 따를 것이라 믿는다. ↓ **4. COMPLIANCE** 준수가 합리적인 행동이 된다. ↓ **5. OBSERVED PERFORMANCE** 실제로 규칙이 반복해서 지켜진다. ↓ **6. TRUST / LEGITIMACY** 제도에 대한 기대가 강화된다. ↓ **7. SHOCK** 위기·위반·정권교체가 온다. ↓ **8. CORRECTION / ADAPTATION** 핵심 기능을 보존하며 수정한다. ↓ 다시: **EXPECTATION RENEWAL** 이다. **제도 복원은 어떻게 가능한가** 단순히: 새 헌법 쓰기. 로 충분하지 않다. 필요한 것은:

- 실제 집행
- 반복된 공정한 사례
- 권력교체
- 예측 가능한 판결
- 제재의 일관성

등이다. 즉: **TRUST IS REBUILT THROUGH PERFORMANCE** 이다. 반대로 붕괴는 더 **빠를 수도 있다** 신뢰는: 오랜 기간 축적. 그러나 결정적 위반 하나가: 규칙이 더 이상 유효하지 않다. 는 강한 신호를 줄 수 있다. 따라서: **TRUST BUILDING SLOW TRUST DESTRUCTION CAN BE FAST** 라는 비대칭이 가능하다. 이것은 **개인관계와도 닮았다** 10년 신뢰. 큰 배신 한 번. 관계 급변. 비유는 제한적이지만: **asymmetric updating** 구조는 유사하다. 그러나 **자본처럼 정확히 측정되는 것은 아니다** 따라서: **TRUST CAPITAL** 은 분석적 비유다. 실제 물리량이 아니다. **HARD LOCK: ANALOGY ≠ IDENTITY** 를 유지한다. 이것은 **개인에게도 그대로 돌아온다** 개인: 습관 문명: 제도 둘 다 과거 선택을 미래 행동의 기본값으로 만든다. 그러나 인간은 Reflection을 통해 습관을 검토할 수 있다. 사회도: **constitutional reform institutional reform** 을 통해 제도를 검토할 수 있다. 즉: **HISTORY CREATES DEFAULTS, NOT DESTINY** 이다. 이 장의 **HARD LOCKS** 규칙 존재 ≠ 제도 작동 제도 작동 ≠ 제도 정당성 신뢰 ≠ 제도 질 높은 신뢰 ≠ 맹목적 순응 경로의존 ≠ 운명론 임계점 ≠ 반드시 예측 가능한 숫자 복원력 ≠ 원상복귀 안정성 ≠ 경직성 효율성 ≠ 복원력 중복 ≠ 단순 낭비 한 번의 위반 ≠ 자동 붕괴 공식 폐지 ≠ 붕괴의 시작점 **REFLEXIVITY ≠ REFLECTION** 이 장의 가장 강한 결론 제도는 사람들이 지키기 때문에 존재하고, 사람들이 지킬 것이라고 믿기 때문에 더 쉽게 지켜지며, 반복된 준수가 다시 그 믿음을 강화한다. 따라서 제도의 실체는: **RULE +**

ENFORCEMENT + EXPECTATION + MEMORY 의 결합이다. 그리고 붕괴는: 규칙문이 사라지는 사건 이 아니라 '다른 사람도 더 이상 이 규칙을 지키지 않을 것'이라는 기대가 행동을 바꾸기 시작하는 과정 일 수 있다. 이제 처음 바둑과 다시 연결한다 바둑에서는: 판 상태 가 돌의 위치로 거의 완전히

보인다. 사회에서는: **제도 상태** 가 보이지 않는 기대까지 포함한다. 따라서 같은 공식 규칙 아래에서도: 전혀 다른 게임이 실제로 진행될 수 있다. 이것이 사회를 바둑보다 훨씬 어렵게 만든다. **사회는 어디에 존재하는가** 공유기대·공통지식·관습·화폐·법·국가·사회적 실재 사회적 존재론은 바로 이런 것들의 존재방식을 연구한다. 돈·기업·정당·국가 같은 것은 단순한 물체도 아니고 단순히 한 사람 머릿속의 생각도 아니다. 사회적 상호작용과 집단적 태도를 통해 성립하면서 실제 행위와 권한을 조직하는 대상이다. **먼저 이상한 것 세 개를 책상 위에 올려놓자** 첫째: 5만 원권. 둘째: 판결문. 셋째: 대통령 임명장. 물리적으로 보면 모두 종이나 전자기록이다. 그런데 전혀 다른 힘을 가진다. 5만 원권을 내면: 물건을 받는다. 판결문이 나오면: 재산이 넘어갈 수 있다. 임명장이 효력을 가지면: 어떤 사람은 국가기관에 명령할 권한을 얻는다. 질문: **이 힘은 종이 어디에 들어 있는가?** 아무 데도 없다. **그렇다면 전부 환상인가** 아니다. 환상이라면: 나는 돈을 믿지 않겠습니다. 라고 선언하면 세금을 안 내도 되고, 나는 법을 믿지 않습니다. 라고 하면 판결에서 자유로워야 한다. 그렇지 않다. 따라서: **SOCIAL CONSTRUCTION ≠ FICTION** 이다. 이것은 이번 노드의 첫 번째 HARD LOCK이다. **화폐를 보면 가장 쉽다** 지폐의 물리적 속성:

- 종이
- 잉크
- 위조방지기술

이다. 하지만 그것만으로는 돈이 아니다. 돈이 기능하려면 다른 사람들이: **이것을 지불수단으로 받아들일 것** 이라는 기대가 중요하다. 화폐철학에서도 현대 화폐를 단순 상품으로 보는 견해와 신용·사회적 관계로 보는 견해가 경쟁하며, 후자의 관점에서는 발행자의 신뢰성과 다른 사람들이 그것을 지급수단으로 받아들일 가능성이 핵심이다. **핵심은 행동 상호의존성이다** 우측통행을 생각하자. 오른쪽이 자연적으로 왼쪽보다 우월한가? 아니다. 중요한 것은: **모두 같은 쪽을 선택하는 것** 이다. 내 최선의 행동은: **다른 사람의 행동에 달려 있다.** 이것이: **COORDINATION PROBLEM** 이다. 관습에 관한 고전적 게임이론 논의에서는 바로 이런 반복되는 조정문제에서 사람들이 특정 균형을 서로가 따를 것이라는 공통지식 아래 따를 때 관습이 성립한다고 본다. 여기서 관습의 본질이 보인다 관습은: 이 행동이 본질적으로 더 좋은 행동이다. 가 아니라 **다른 사람도 이 행동을 할 것이라고 기대하기 때문에 내가 하는 행동** 일 수 있다. 즉: 이라는 강한 후보가 생긴다. 그런데 현실 인간이 무한 재귀를 계산할까 아니다. 우리가 실제로: 그가 알고 내가 알고 그가 내가 아는 것을 알고... 를 무한히 계산하는 것은 아니다. 공통지식은: **실제 뇌 안의 무한 연산** 이라기보다 **사회적 조정상태를 모델링하는 이상화된 개념** 으로 보는 편이 안전하다. **HARD LOCK: COMMON KNOWLEDGE ≠ INFINITE CONSCIOUS RECURSION** 이다. 여기서 사회적 존재론의 핵심 문제가 나온다 돈. 법. 대학. 기업. 국가. 이런 것은: **개인 한 명 안에 존재하지 않는다.** 그렇다고: **공중에 떠 있는 독립 물체도 아니다.** 그것들은: **여러 사람·규칙·물체·기대·역할 사이의 조직된 관계** 로 존재한다. 아주 중요한 구분 **COLLECTIVE INTENTIONALITY** 는: 여러 사람이 어떤 대상·목표·믿음에 공동으로 향하는 능력 을 뜻하는 철학적 범주다. 예:

- 함께 아이를 돌봄
- 정당운동
- 공동프로젝트
- 공동인정

등이다. 공유의도·공유믿음·집단적 수용 등이 언어·상징·제도와 사회적 지위를 구성하는 데 중요한 역할을 한다는 것이 이 연구전통의 핵심 주장 가운데 하나다. **그래도 집단행위는 실제일 수 있다** 기업이 계약을 체결한다. 국가가 조약을 체결한다. 대학이 학위를 수여한다. 한 명의 개인행위로 환원해 설명하기 어려운: **조직적 의사결정 구조** 가 있다. 따라서: **GROUP ACTION IS NOT NECESSARILY FICTION** 이다. 하지만 '다들 믿으니까 권력'도 부족하다 무장집단이 있다. 아무도 그 권한을 정당하다고 인정하지 않더라도: **물리적 강제력** 으로 행동을 바꿀 수 있다. 따라서: **POWER ≠ LEGITIMACY POWER ≠ SHARED BELIEF ALONE** 이다. 사회적 실재에는:

- 물리적 능력
- 제도적 권한
- 사회적 인정

이 서로 다른 비중으로 결합될 수 있다. **프로젝트용 국가 최소모델 TERRITORY** 일정한 공간. **POPULATION** 구성원. **ORGANIZATIONAL CAPACITY** 행정·조세·집행능력. **RULES / ROLES** 권한구조. **MUTUAL RECOGNITION** 시민·기관·외부 행위자가 그 구조를 국가로 취급. 이런 요소의 결합으로 보는 편이 안전하다. **아주 중요한 원리 DEPENDENT REALITY CAN STILL BE REAL** 이다. 예: 온도. 분자운동에 의존한다. 그렇다고 온도가 가짜인가? 아니다. 사회적 실재도 인간 관계에 의존한다고: **허구** 가 되는 것은 아니다. 물론 온도와 사회제도는 동일한 존재방식을 갖는다는 뜻은 아니다. **암호화폐가 재미있는 이유** 질문: 국가가 발행하지 않으면 돈일 수 없는가? 바로 사회적 존재론 문제다. 어떤 암호자산이:

- 교환수단
- 가치저장
- 가격표시

기능을 실제로 수행한다면 일부 화폐적 성격을 가질 수 있다. 하지만: **기술적 토큰 존재와 사회적 화폐기능** 은 동일하지 않다. 화폐철학에서도 바로 이런 이유로 암호화폐가 “무엇이 돈인가”라는 존재론적 질문을 다시 강하게 만든다. **사회적 실재에는 '실효성'이 중요하다** 따라서 사회적 대상 X의 존재를: **사람들이 X라고 부른다** 만으로 판단하면 부족하다. 봐야 한다. 그 지위가 실제 행동과 권리·의무를 체계적으로 바꾸는가? 즉: **SOCIAL STATUS → BEHAVIORAL CONSEQUENCE** 인터페이스가 필요하다. **나는 세금을 싫어할 수 있다** 하지만 안다. 세금을 내지 않으면 이런 일이 벌어진다. 그리고: 세무당국에게 징수권한이 있다고 사회가 운영되고 있다. 고 행동한다. 즉: **NORMATIVE SUPPORT** 가 없더라도 **PRACTICAL RECOGNITION** 이 있을 수 있다. **이제 바둑으로 돌아가자** 흑돌 네 개가 어떤 구조를 만든다. 우리는 말한다. 두터움이 있다. 집 모양이다. 약한 돌이다. 질문: **두터움은 어느 돌 안에 들어 있는가?** 없다. 두터움은: 돌들 사이 관계 + 주변 빈 공간 + 미래 전투가능성 에서 나온다. 즉: **RELATIONAL PROPERTY** 다. **사회적 실재와 완전히 같다는 뜻은 아니다** 바둑의 두터움은: 전략적 관계구조. 돈은: 사회적·제도적 관계구조. 둘은 다르다. 그러나 유용한 공통점: **중요한 속성이 구성요소 하나에 들어 있지 않고 관계망에서 나타날 수 있다.** 이다. **HARD LOCK EMERGENCE ≠ MAGIC** 이다. “창발했다.” 라고 말하는 것은 설명의 끝이 아니다. 반드시 물어야 한다. **어떤 미시 상호작용이 어떤 전달경로를 통해 거시패턴을 안정화했는가?** 즉: **MICRO → INTERFACE → MACRO** 가 필요하다. 예 — 언어 개인이 단어를 사용. ↓ 다른 사람이 이해. ↓ 반복. ↓ 기대 형성. ↓ 교정과 교육. ↓ **안정된 언어규칙.**

언어관습도 조정균형과 공통지식의 대표적 사례로 모델링되어 왔다. 여기에 S6을 추가할 수 있다 ENFORCEMENT / CAPACITY 특히 법·국가 같은 제도에서는: 규칙 위반 시 실제 대응할 능력이 있는가?가 중요하다. 따라서 강한 제도에는: S6 — ENFORCEMENT 가 추가된다. 소문을 검사하면 사람들이 믿는다. 하지만:

- 안정된 역할
- 규칙
- 지속성

이 없을 수 있다. 따라서: SHARED BELIEF 지만 INSTITUTION 은 아닐 수 있다. 좋은 경계다. 사회적 실재도 층위가 있다 LEVEL 1 — SHARED PATTERN 비슷한 행동. LEVEL 2 — CONVENTION 상호기대에 의해 유지되는 조정. LEVEL 3 — NORM 해야 한다/하면 안 된다는 기대와 제재. LEVEL 4 — INSTITUTION 역할·규칙·지속적 조직구조. LEVEL 5 — ORGANIZED AUTHORITY 법·국가처럼 집행능력을 가진 체계. 이것은 절대적 분류가 아니라 프로젝트용 작동지도다. 규범은 왜 강인가 누군가 위반한다. 다른 사람이: 이상하다. 뿐 아니라 그러면 안 된다. 고 반응한다. 즉 규범은: 2차 반응 을 만든다. 행동에 대한 행동. 이것이 규범을 자기강화할 수 있게 한다. 법은 규범과 무엇이 다른가 모든 사회규범이 법은 아니다. 예: 식사예절. 법에는 추가로:

- 공식 권한
- 절차
- 집행기관

이 있다. 즉: LAW $\approx$ NORM + AUTHORIZED PROCEDURE + ENFORCEMENT 정도의 프로젝트 지도는 가능하다. 단, 법철학의 완전한 정의는 아니다. 이것은 '분산된 실재'다 뇌의 기억도 한 뉴런에 들어 있지 않을 수 있다. 회사의 존재도 한 직원에 들어 있지 않다. 통화도 한 지폐에 들어 있지 않다. 즉: DISTRIBUTED REALIZATION 이라는 표현이 유용하다. 하지만: 사회: 뇌 라는 동일성 주장은 아니다. 우선 세 층을 분리한다 COLLECTIVE BEHAVIOR 사람들이 함께 행동. COLLECTIVE AGENCY 조직된 절차를 통해 집단 이름으로 결정·행동. COLLECTIVE CONSCIOUSNESS 집단 자체가 하나의 주관적 경험을 가짐. 앞의 둘은 다양한 의미에서 충분히 논의 가능하다. 세 번째는: 전혀 다른 강한 주장 이다. 사회적 실재의 가장 흥미로운 특징 일단 만들어지면: 그것이 우리를 다시 만든다. 화폐를 만든다. ↓ 화폐 때문에:

- 직업
- 가격
- 부
- 선택

이 달라진다. 학교를 만든다. ↓ 학교가:

- 지식
- 신분
- 인생경로

를 바꾼다. 즉: HUMAN → INSTITUTION → HUMAN 이라는 페루프가 있다. 이것을 앞 장 자유의지와 연결하면 개인은: 사회적 선택환경 속에서 선택한다. 언어. 돈. 직업. 법. 가족역할. 이미 존재한다. 그러나 Reflection을 통해: 그 규범과 역할 자체를 검토 할 수도 있다. 즉: Institution ↓ Self formation ↓ Reflection ↓ Institutional change 이다. 여기서 '문화'의 위치도 보인다 문화는: 하나의 규칙집 이 아니다.

- 언어
- 의미
- 관습
- 규범
- 상징
- 기억

이 여러 방식으로 전달되는 분산구조다. 따라서 문화는: SOCIAL MEMORY + EXPECTATION ENVIRONMENT로 볼 수 있는 측면이 있다. 단, 문화 전체를 이 정의로 환원하지 않는다. 문명이라는 말이 다시 정확해진다 앞 장에서 제안했다. 문명은 개인의 시행착오가 개인과 함께 사라지지 않도록 외부에 저장하는 누적학습체계 라고. 앞 장가 이를 보강한다. 저장은 단지: 정보를 기록 하는 것이 아니다. 그 정보를 역할·규칙·기대 속에 넣어 다음 세대의 기본 행동환경으로 만드는 것 이다. 이것이 사회가 물질세계에 다시 들어오는 인터페이스다 추상적 규칙: 소유권. ↓ 법적 기록. ↓ 은행 담보. ↓ 대출. ↓ 건물 건설. 즉: SOCIAL ABSTRACTION → PHYSICAL CONSEQUENCE 가 된다. 사회적 실재는 추상적이지만: 물질 흐름을 조직한다. MASTER MODEL v1.9 SOCIAL REALITY LOOP PHYSICAL / HUMAN BASE 사람·사물·기술. ↓ SHARED REPRESENTATION “이것은 돈이다.” “이 사람은 판사다.” ↓ MUTUAL EXPECTATION 다른 사람도 그 지위에 맞게 행동할 것이다. ↓ RULE / ROLE 어떤 권리·의무·행동이 연결되는가. ↓ ENFORCEMENT / REINFORCEMENT 규칙 준수와 위반에 결과가 생긴다. ↓ BEHAVIORAL REGULARITY 반복되는 사회행동. ↓ EXTERNAL RETENTION 문서·교육·조직·관행에 저장. ↓ NEXT GENERATION EXPECTATION 다음 사람은 이미 만들어진 세계 안으로 들어온다. ↓ 다시: SHARED REPRESENTATION 이다. HARD LOCKS v1.9 사회적으로 구성됨 ≠ 허구 공유믿음 ≠ 제도 공통지식 ≠ 실제 무한재귀 사고 인정 ≠ 찬성 사회적 존재 ≠ 정당성 집단행동 ≠ 집단의식

집단행위자 ≠ 거대한 개인 권력 ≠ 물리적 강제력만 법조문 ≠ 법의 실효적 작동 전체 관습 ≠ 법 창발 ≠ 설명 관계적 속성 ≠ 가짜 속성 이 장의 가장 강한 결론 사회는 사람들의 머릿속에만 있지 않다. 사람들 사이의 기대, 물질적 기록, 반복되는 역할, 집행능력과 행동관계 속에 분산되어 존재한다. 따라서: 돈은: 종이가 아니다. 하지만 종이와 인간 없이 존재하는 플라톤적 대상도 아니다. 법은: 문장이 아니다. 하지만 사람들의 단순 의견도 아니다. 국가는: 한 사람이 아니다. 하지만 단순한 허구적 이름도 아니다. 모두: RELATIONALLY REAL 이다. 그리고 처음 바둑과 아주 흥미롭게 연결된다 바둑판에서: 두터움은 돌 하나 안에 없다. 돌들의 관계와 미래 가능성에서 존재한다. 사회에서: 권력 도 개인 한 명 안에 전부 들어 있지 않을 수 있다.

- 명령할 지위
- 명령을 집행할 사람
- 복종 기대

- 법
- 기록
- 강제수단

사이에서 존재한다. 즉 우리가 세계를 이해하려면: **THINGS** 뿐 아니라 **RELATIONS** 을 봐야 한다. **이것이 프로젝트를 다시 ‘우주’ 쪽으로 돌릴 중요한 지점이다** 지금까지 계속 발견했다. 바둑: 돌보다 관계. 생명: 분자보다 조절망. 의식: 신호 하나보다 통합된 과정. 사회: 개인 하나보다 관계망. 그러면 매우 위험하지만 중요한 질문이 열린다. **세계의 근본은 ‘것’인가 ‘관계’인가?**

제4부

세계를 안다는 것

관계·과정·정보·모델을 거쳐 참과 의미의 서로 다른 판정구조를 분리한다.

10장. 관계·과정·자율성

계속 변하는 시스템을 같은 존재라고 부르고 자율적이라고 말할 근거는 무엇인가?

어떤 존재의 정체성은 고정된 재료보다 관계와 조직, 경계를 유지·갱신하는 과정에 있을 수 있다. 그러나 이를 모든 존재의 단일 형이상학으로 확대해서는 안 된다. 사물보다 관계가 더 근본적일 수 있는가 바둑의 형세 · 생명의 조직 · 뇌의 네트워크 · 사회제도 · 물리학의 관계적 기술 관계적 속성에서 관계적 존재론으로 넘어갈 수 있는가 가장 안전한 결론은: **THING-ONLY EXPLANATION IS OFTEN INSUFFICIENT.** 이지 **ONLY RELATIONS EXIST.** 가 아니다. 과학철학의 구조적 실재론도 바로 이 경계에서 여러 버전으로 갈린다. 우리가 세계의 구조를 가장 안정적으로 안다는 인식론적 주장과, 구조나 관계가 객체보다 존재론적으로 더 근본적이라는 강한 주장은 동일하지 않다. **먼저 바둑돌 하나를 보자** 흑돌 하나. 그 자체로 알 수 있는 것:

- 검다.
- 교차점에 놓여 있다.

그러나 바둑에서 정말 중요한 질문: 이 돌은 강한가? 약한가? 두터운가? 죽었는가? 중요한가? 이것들은 돌 하나만 보면 알 수 없다. 주변 돌. 빈 공간. 연결. 절단. 수순. 전체 형세. 가 필요하다. 즉: **PROPERTY OF STONE ≠ PROPERTY OF POSITION** 이다. 돌을 하나도 바꾸지 않고 관계만 바꿀 수 있다 흑돌 네 개. 같은 돌 네 개. 하지만 배치가 달라지면:

- 강한 모양
- 끊긴 모양
- 빈삼각
- 두터운 모양

이 된다. 물질: 동일. 관계: 다름. 결과: 기능 다름. 따라서: **COMPOSITION ALONE DOES NOT DETERMINE FUNCTION** 이라는 첫 번째 원리가 나온다. 두 번째 구분 **STRUCTURAL EXPLANATION** 개별 구성요소의 속성보다: **배치·연결구조** 가 현상을 설명하는 데 중요하다. 예: 네트워크. 같은 노드도: 누구와 연결되는가에 따라 전체 동역학이 바뀐다. 2024년 시스템생물학 리뷰에서도 네트워크가 지원할 수 있는 동적 상태는 단순 구성요소 목록이 아니라 **네트워크 위상과 상호작용 논리**에 의해 결정되는 문제로 다뤄진다. 이 역시: **STRUCTURE MATTERS** 를 보여준다. 하지만 아직: **ONLY STRUCTURE EXISTS** 는 아니다. **생명을 보자** 세포 안에:

- DNA
- 단백질
- 지질
- 당
- 이온

이 있다. 물질 목록을 모두 안다고 하자. 그것만으로: **살아 있음** 을 설명할 수 있을까? 충분하지 않다. 중요한 것은:

- 어떤 분자가 어떤 반응을 촉진하는가
- 막이 무엇을 분리하는가
- 유전자 발현이 어떻게 조절되는가
- 세포들이 어떻게 신호를 주고받는가

이다. **이것이 네트워크 과학의 교훈이다** 네트워크는: **NODE + EDGE** 다. 노드만 보면 부족하다. 하지만 edge만 있는 것도 아니다. 그리고 더 중요하게: **같은 topology** 라고 같은 시스템도 아니다. 뉴런 연결망. 도로망. 친구관계망. 모두 graph로 표현할 수 있다. 그러나: **EDGE MEANING** 이 다르다. **이것이 이 프로젝트의 핵심 방어선이다** 바둑도 network. 뇌도 network. 사회도 network. 우주도 network. 따라서: 다 같은 원리다. NO. 수학적 표현이 동일하면: **비교할 수 있는 질문** 은 생긴다. 하지만: **동일성** 은 증명되지 않는다. **뇌에서도 마찬가지다** 뉴런 하나를 분석한다. 막전위. 수용체. 축삭. 중요하다. 하지만:

- 기억
- 주의
- 행동정책

같은 현상은 여러 뉴런 사이의 활동패턴과 회로 수준 상호작용을 요구한다. 따라서: **NEURON-ONLY EXPLANATION** 은 충분하지 않을 수 있다. **하지만 이것도 아직 '관계만 존재'가 아니다** 일반상대론에는: **metric field** 가 있다. 물질장도 있다. 사건과 시공간구조를 어떻게 존재론적으로 해석할지는 별도 철학문제다. 따라서: **GENERAL RELATIVITY PROVES RELATION-FIRST ONTOLOGY** 라고 말하면 안 된다. 즉 **물리학도 우리에게 이런 교훈을 준다 REPRESENTATION ≠ REALITY** 이다. 수학적 변수 하나하나가: **독립적 실재** 를 가리킨다고 자동으로 생각하면 안 된다. 어떤 것은: **표현의 중복** 일 수 있다. 관측 가능한 것은 불변량이나 관계구조에 더 직접적으로 연결될 수 있다. **RQM은 정말 강한 관계주의다** 아주 단순화하면: 고전적 그림: **System S has property X**. RQM: **S has value X relative to system O in an interaction**. 즉 물리적 값이: **관계적 사건** 으로 나타난다. RQM의 핵심 설명도 물리량이 상호작용에서 값을 얻고 그 값이 다른 시스템에 상대적이라는 것이다. **이제 Structural Realism** 과학의 역사를 보면: 이론이 바뀐다. 전자에 대한 그림도 바뀐다. 빛에 대한 그림도 바뀐다. 그런데: **수학적 구조와 관계 일부** 는 후속 이론에 살아남는 경우가 있다. 그래서: 과학이 가장 잘 포착하는 것은 객체의 내재적 본성보다 구조일 수 있다. 는 생각이 나온다. **전자와 왕을 동일시하면 안 된다** 왕: **institution-dependent entity**. 전자: **fundamental physical entity candidate**. 관계가 중요하다는 한 문장으로 둘을 같은 종류로 만들면: **category error** 다. 따라서: **RELATIONALITY HAS DOMAIN-SPECIFIC MEANINGS**

를 잠근다. **관계적 설명의 작동조건을 만들자** **R1 — DEPENDENCE** 대상의 중요한 속성이 다른 대상 또는 구조에 의존하는가? **R2 — COUNTERFACTUAL SENSITIVITY** 구성요소는 그대로 두고 관계만 바꾸면 결과가 변하는가? **R3 — MECHANISM** 그 관계가 결과를 바꾸는 전달경로가 있는가? **R4 — SCALE** 어느 수준에서 관계가 설명력을 갖는가? **R5 — NON-REDUNDANCY** 구성요소 속성만으로는 같은 설명을 얻기 어려운가? 이 조건을 많이 충족할수록: **RELATIONAL EXPLANATION EARNS**

ITS NAME 이다. 사례 — 전자 여기서는 조심한다. 전자 성질의 물리적 기술에는 다른 장·시스템과의 관계가 중요하다. 게이지 이론과 일부 양자해석은 물리적 내용의 관계적 성격을 강하게 강조한다. 하지만: 전자가 오직 관계로만 존재한다 는 결론은: **PHILOSOPHICALLY CONTESTED**. 이다. 따라서 **DOMAIN VERDICT GO** 바둑: 관계적 설명 강함. **GO** 생물·뇌: 구성요소 + 관계 + 동역학 필요. **GO** 사회: 사회적 종류 상당수가 본질적으로 관계적·제도적. **PARTIAL** 물리학: 많은 핵심 양과 이론에서 관계·불변구조가 중요. **HOLD** 우주의 최종 존재론: **관계가 객체보다 근본이라는 결론**. 이제 '관계'와 '정보'를 연결해보고 싶은 유혹이 생긴다 관계가 있다. 그 관계가 차이를 만든다. 그러면: 관계 = 정보? 라고 하고 싶어진다. 하지만: **NO**. 정보 역시 여러 의미가 있다. Shannon information. semantic information. biological information. knowledge. 모두 다르다. 앞 장에서 이미: **INFORMATION ≠ KNOWLEDGE** 를 잡았다. **관계와 상호작용도 다르다** 서울은 부산 북쪽에 있다. 관계. 하지만 둘이 서로 물리적으로 상호작용한다는 뜻은 아니다. 따라서: **RELATION** 은: **INTERACTION**

보다 넓다. 그리고: **INTERACTION** 은 보통 실제 상태변화를 일으키는 동역학적 연결이다. 그러면 '우주의 이치'라는 처음 질문에 어떤 답을 줄 수 있는가 아직: 모든 것은 관계다. 라고 할 수 없다. 하지만 다음은 상당히 강하다. 세계의 많은 중요한 현상은 개별 사물의 속성을 따로 더하는 것만으로 설명되지 않으며, 구성요소들이 어떤 관계와 상호작용 구조 안에 놓여 있는지를 함께 알아야 한다. 이것은 존재론 선언이 아니라: 설명원리 다. 그리고 사회도 자연과 동일한 네트워크가 아니다 사회에는:

- 의미
- 규범
- 권리
- Reflection

이 있다. 따라서 물리학적 network model 하나로 사회 전체를 설명할 수 없다. 하지만: **관계적 구조가 행위결과를 바꾼다** 는 비교질문은 유효하다. 이것은 새로운 문제다 **IDENTITY THROUGH CHANGE** 무엇이 계속 바뀌는데: 왜 같은 것이라고 부르는가? 배: 널빤지 교체. 사람: 세포·기억 변화. 회사: 직원 전부 교체. 국가: 세대 전체 교체. 그럼: 무엇이 동일성을 보존하는가? 이다. 이것은 관계론보다 더 검증가치가 높다 왜냐하면 지금까지 거의 모든 프로젝트에서: **TIME** 이 중요하게 반복됐기 때문이다.

- 학습
- 진화
- 자기서사
- 제도
- 경로의존
- 신뢰

모두: 시간 없이는 존재할 수 없는 구조 다. 따라서 다음에는: 관계가 아니라 과정 을 검증해야 한다. **HARD LOCKS v2.0** 관계적 속성 ≠ 관계적 존재론 구조적 설명 ≠ 구조만 존재 네트워크 ≠ 메커니즘 전체 같은 그래프 ≠ 같은 현상 관계 ≠ 상호작용 상호작용 ≠ 인과 전체 RQM ≠ 양자역학 전체의 유일한 해석 RQM ≠ 의식이 현실을 만든다는 주장 상대성 ≠ 모든 것이 주관적 구조적 실재론 ≠ 확정된 물리학

결론 창발 ≠ 존재론적 근본성 보편적 적용가능성 ≠ 높은 설명력 이 장의 가장 강한 결론 복잡한 세계를 이해할 때 '무엇이 있는가'만큼 '무엇이 무엇과 어떤 방식으로 연결되어 있고, 그 연결이 시간에 따라 무엇을 가능하게 하는가'를 물어야 한다. 그러나: 관계를 봐야 한다 에서 관계만 존재한다 로 넘어가지는 않는다. 현재 가장 강한 통합은: THINGS + RELATIONS + DYNAMICS 이다. 그리고 처음 질문에 대한 중간답 “바둑의 알고리즘과 세상의 모든 이치가 연결되는가?” 현재 답은: 동일한 알고리즘으로는 연결되지 않는다. 하지만: 좋은 수를 이해하려면 돌보다 형세를 봐야 했듯, 많은 현실 시스템에서도 구성요소보다 구성요소 사이의 관계와 그 관계가 만드는 미래 가능성까지 봐야 한다. 이것은 살아남는다. 다음 질문은 이제 더 깊다 사물이든 관계든: 시간 속에서 계속 변한다. 그런데도:

- 나는 계속 나인가?
- 생명체는 계속 같은 생명체인가?
- 회사는 같은 회사인가?
- 국가는 같은 국가인가?

즉: 지속되는 것은 물질인가, 구조인가, 과정인가? 이 문제가 열린다. 모든 것이 바뀌는데 왜 같은 것이라고 부르는가 과정 · 동일성 · 자기유지 · 기억 · 인과적 연속성 · 테세우스의 배 생명 · 자아 · 조직 · 국가의 지속성 과정철학은 지속하는 사물을 정적인 물질덩어리보다 시간 속에서 안정적으로 재생산되는 동적 조직으로 이해하려 한다. 특히 생물철학에서는 유기체의 지속성과 개체성을 인과적으로 유지되는 과정으로 설명하려는 접근이 진지하게 발전해 왔다. 그러나 이것은 여전히 하나의 철학적 프로그램이지 모든 존재에 대해 확정된 결론은 아니다. 테세우스의 배 배 한 척이 있다. 널빤지가 썩었다. 한 장 교체. 또 한 장. 또 한 장. 오랜 시간이 지나: 처음 널빤지가 하나도 남지 않았다. 질문: 같은 배인가? 직관적으로는: 그렇다. 라고 말할 수 있다. 왜? 한꺼번에 새 배를 만든 것이 아니라: 계속 운항하던 배를 조금씩 수리했기 때문이다. 그러면 첫 번째 후보가 나온다. IDENTITY MAY TRACK CONTINUITY, NOT MATERIAL SAMENESS. 여기서 두 기준이 충돌한다 MATERIAL CONTINUITY 원래 물질이 남아 있는가? → B에게 유리. CAUSAL / FUNCTIONAL CONTINUITY 원래 배의 활동과 역사가 끊기지 않았는가? → A에게 유리. 즉: IDENTITY CRITERIA CAN CONFLICT. 이것이 퍼즐의 핵심이다. 그러나 생명에 가면 과정적 설명이 훨씬 매력적이다 당신 몸의 물질은 계속 바뀐다. 먹는다. 숨 쉰다. 배설한다. 세포가 죽고 새로 생긴다. 분자가 교체된다. 그런데도 우리는: 어제와 오늘 같은 유기체. 라고 말한다. 왜? 같은 원자가 남아서 는 아니다. 그래서 생명체의 정체성에는 자기유지가 중요하다 외부에서 누군가 계속: 이 형태로 유지하세요. 라고 조립하는 것이 아니다. 유기체 내부의 여러 과정이: 그 유기체의 존속조건을 계속 재생산 한다. 예:

- 막 유지
- 에너지 대사
- 단백질 교체
- 손상복구
- 체온조절

즉: SELF-MAINTENANCE. 예를 들어 산호 산호군체는: 하나인가? 많은 개체인가? 군류의 군사체는? 점군류는? 홀로바이온트는? 어디서: 하나가 끝나고 여러 개가 시작되는가? 생명은 우리 직관보다 경계가 흐릴 수 있다. 그래서 생물에서는 아주 강한 식이 나온다 PERSISTENCE ≠ MATERIAL CONSTANCY. 더 좋은 후보: 즉: 물질이 바뀌어도 조직이 자기 자신을 인과적으로 이어간다.

이건 매우 중요한 전환이다 THING-FIRST: 지금 무엇인가? PROCESS-FIRST: 어떻게 여기까지 왔는가? 즉 현재의 정체성을 이해하려면: **HISTORY BECOMES CONSTITUTIVE**. 과거가 단순 배경정보가 아니라 현재가 무엇인지의 일부가 된다. **생명도 같다** 체온이 떨어졌다. 조절한다. 상처났다. 복구한다. 환경 바뀔. 행동 바뀔. 즉: **유지하기 위해 변한다**. 이것이 이 장의 가장 중요한 역설 중 하나다. **CHANGE CAN BE THE MECHANISM OF PERSISTENCE**. 하지만 또 동일시하면 안 된다 자전거. 유기체. 국가. 모두 “동적 안정”이라고 부를 수 있지만: **유지 메커니즘은 전혀 다르다**. 따라서: NOT UNIVERSAL MECHANISM. 이다. **회사 동일성은 생물 동일성과 다르다** 기업에서는:

- 법인 등록
- 계약관계
- 자산·부채
- 조직권한
- 이름과 기록

같은: **institutional continuity** 가 중요하다. 즉 사람의 동일성과 달리: **법적 규칙이 동일성을 구성** 할 수 있다. **회사를 분할하면?** 회사 A를: A1 + A2 로 분할. 어느 쪽이 원래 A인가? 법은: **승계 규칙** 을 정할 수 있다. 즉 사회적 대상의 동일성은 일부: **CONVENTIONALLY / LEGALLY SPECIFIED** 될 수 있다. 생물개체와 큰 차이다. 그리고 **정치적 동일성은 논쟁 대상이 되기도 한다** “우리가 진짜 그 국가의 계승자다.” “이 정권은 이전 체제와 단절됐다.” 이런 주장은 단순 사실기술이 아니라: **정치적·법적 이해관계** 를 포함할 수 있다. 따라서: **IDENTITY CLAIM CAN ITSELF BE AN ACT OF POWER**. 이다. 여기서 전체 프로젝트에 필요한 구분이 나온다 **TYPE 1 — MATERIAL CONTINUITY** 같은 물질이 얼마나 유지되는가. **TYPE 2 — STRUCTURAL CONTINUITY** 배치·조직이 얼마나 유지되는가. **TYPE 3 — CAUSAL CONTINUITY** 이전 상태가 적절한 인과경로를 통해 다음 상태를 만드는가. **TYPE 4 — FUNCTIONAL CONTINUITY** 핵심 기능이 유지되는가. **TYPE 5 — INFORMATIONAL / MEMORY CONTINUITY** 과거 상태의 정보가 보존되는가. **TYPE 6 — INSTITUTIONAL CONTINUITY** 규칙·권리·역할·승계가 유지되는가. 하나의 동일성 기준으로 모든 분야를 설명하면 안 된다. **사례별로 넣어보면** • 대상 — 강한 지속성 후보 • 바위 — 물질·공간적 연속성 • 불꽃 — 과정적·인과적 연속성 • 유기체 — 자기유지 조직 + 인과적 발달 • 사람 — 생물학적 + 심리적 인과연속성 논쟁 • 회사 — 법적·조직적 승계 • 국가 — 제도·영토·국제적·역사적 연속성의 결합 즉: **ONE THEORY OF IDENTITY FOR EVERYTHING** 은 현재 거부한다.

그러나 ‘모든 것이 과정’은 아직 별도 주장이다 둘도 원자수준에서는 변화한다. 별도 과정이다. 시공간도 동적이다. 그래서: 결국 모든 것은 과정이다. 라고 말하고 싶어진다. 하지만 다시: **너무 보편적이면 구별력을 잃는다**. 바위 P1 시간에 존재. P2 내부변화가 정체성의 핵심인가? 상대적으로 약함. P3 있음. P4 적극적 재생산 약함. P5 유지과정을 ‘정지’한다고 바위가 바로 사라지는 것도 아님. 따라서: **PROCESS DESCRIPTION POSSIBLE, BUT PROCESS EXPLANATION LESS NECESSARY**. 좋은 구별이다. 여기서 **생명과 VALUE가 다시 연결된다** 왜 유기체가:

- 먹고
- 도망가고
- 회복하는가?

자기 조직을 유지해야 하기 때문이다. 즉 앞 장의: **VIABILITY** 가 과정론으로 연결된다. 살아 있는 존재에게: **VALUE** 가 생기는 가장 낮은 층은: **자신의 지속과 관련된 상태차이** 일 수 있다. 그래서 흥미로운 연결고리가 생긴다 **SELF-MAINTENANCE** ↓ **VIABILITY BOUNDARY** ↓ **STATES MATTER DIFFERENTLY** ↓ **BIOLOGICAL VALUE** 이다. 이걸 상당히 강한 프로젝트 연결이다. 다만: 자기유지: 의식적 목적 은 아니다. **HARD LOCK SELF-MAINTENANCE** ≠ **DESIRE TO SURVIVE**. 이다. 유기체가 자기유지하도록 조직됐다는 것과: 살고 싶어 한다. 는 전혀 다른 주장이다. 극단적 사례가 경계를 보여준다 심한 기억상실. 질문: 같은 인간 유기체인가? 대체로 YES. 같은 법적 사람인가? 보통 YES. 같은 심리적 인격인가? 복잡해짐. 즉 동일성은: **DOMAIN-DEPENDENT**. 이것은 앞 장의 미래 자기와 연결된다 현재의 내가: 저축한다. 운동한다. 공부한다. 왜? 미래의 나를 현재 의사결정 안에 넣기 때문이다. 만약 미래의 나를: 완전히 다른 사람. 처럼 느끼면? 장기투자 감소 가능. 즉 자기 동일성의 느낌은: **intertemporal decision** 과 연결된다. 개인 기억

- 신경계
- 몸
- 기록

등. 제도 기억:

- 문서
- 역할
- 법
- 관행
- 건물
- 데이터

즉 문명은: **MEMORY OUTSIDE THE ORGANISM** 을 만들어낸다. 앞 장의 핵심이 다시 강화된다. 이것이 과정론의 한계도 보여준다 “인과적 연결이 있으니 같은 과정.” 이라고만 하면: 거의 모든 역사를 하나로 묶을 수 있다. 예: 로마 → 유럽 → EU? 인과관계 있음. 그렇다고: 같은 정치적 개체 는 아니다. 따라서: **CAUSAL CONTINUITY** ≠ **IDENTITY BY ITSELF**. 이다. 하지만 **invariant**: 완전 불변도 아니다 체온: 정확히 36.5000도 고정? 아니다. 범위. 조직도 변화. 따라서: **INVARIANT CAN BE A PATTERN OF CONTROLLED VARIATION**. 이 문장이 중요하다. 생명에서도 세포 교체: 허용. 체중 변화: 허용. 상처: 복구. 하지만 특정 임계점을 넘으면: 생명 유지 실패. 즉: **IDENTITY AND VIABILITY HAVE BOUNDARIES**. 이제 **PROCESS-FIRST**의 가장 강한 형태 사물을: 변화하지 않는 핵 + 부수적 변화 로 보는 대신 **변화의 흐름 속에서 특정 조직을 계속 재생산하는 패턴** 으로 보는 것이다. 이것이 생명·자아·제도에서는 상당히 강력하다. 생명에서 가장 선명하다 대사 중단. 조절 중단. 복구 중단. 분자는 잠시 남는다. 몸 모양도 남는다. 하지만: 살아 있는 유기체로서의 지속성 은 끝난다. 따라서 생명에 관해서는: **TO BE IS, IN PART, TO KEEP GOING**. 이라는 과정적 직관이 상당히 강하다. 이 장 **HARD LOCKS** 물질교체 ≠ 동일성 상실 같은 모양 ≠ 같은 개체 인과연속성 ≠ 자동 동일성 기억 ≠ 개인정체성 전체 자기서사 ≠ 수적 동일성 변화 ≠ 정체성 상실 안정성 ≠ 무변화 자기유지 ≠ 생존 욕망 생물 개체성 ≠ 하나의 합의된 정의 회사 동일성 ≠ 생물 동일성 국가 동일성 ≠ 자연적 단일 기준 과정적 설명 성공 ≠ 과정만 존재한다는 존재론 이 장의 최종 판정 복잡한 지속체는 동일한 물질을

계속 보유해서 같은 것이기보다, 과거 상태에서 현재 상태로 이어지는 적절한 인과적 연속성과 조직적 자기유지·기억·승계 구조를 통해 지속할 수 있다. 그러나 ‘적절한’ 것이 무엇인지는 대상마다 다르다. 따라서: 동일성은 하나의 보편적 물질기준이 아니라 영역별 지속성 메커니즘을 판별해야 하는 문제다. 그리고 아주 큰 질문이 열린다 유기체는: 자기 자신을 계속 만들어낸다. 자아는: 자기 자신을 계속 수정한다. 제도는: 구성원들이 계속 재생산한다. 그렇다면: 자기유지와 자기생성이 단순한 결과가 아니라 시스템을 하나의 ‘개체’로 만드는 핵심인가? 이 질문이 나온다. 이번에는 정의를 아주 좁게 잡아야 합니다. 자기조직화는 패턴이 스스로 생기는 것, 자기유지는 그 패턴을 계속 유지하는 것, autopoiesis는 자기 구성요소와 경계를 생산하는 네트워크, agency는 그 위에서 환경에 따라 선택적으로 행동을 바꾸는 능력으로 분리하겠습니다. 최근 문헌에서도 이 경계들은 여전히 논쟁적이며, 특히 agency를 생물 일반의 속성으로 얼마나 넓게 잡을지는 열려 있습니다. 시스템은 언제 ‘자기 자신’을 만들어내는가 자기조직화 · 자기유지 · autopoiesis · 경계 · 대사 · 자율성 · agency 가장 안전한 층위는: SELF-ORGANIZATION ↓ SELF-MAINTENANCE ↓ SELF-PRODUCTION / AUTOPOIESIS ↓ AUTONOMOUS REGULATION ↓ AGENCY 이다. 그러나 이것은 필연적 진화계단이 아니라 기능적으로 더 강한 조건이 추가되는 방향이다. 먼저 눈송이를 보자 과냉각된 수증기가 있다. 조건이 맞으면: 정교한 결정구조 가 형성된다. 누가 설계했는가? 아무도 없다. 국소적 물리법칙과 상호작용만으로: 패턴이 생겼다. 이것은: SELF-ORGANIZATION 의 후보 사례다.

그런데 눈송이는: 깨지면 스스로 수리하지. 라고 하지 않는다. 자기 부품을 계속 생산하지도 않는다. 따라서: SELF-ORGANIZATION ≠ SELF-MAINTENANCE. 자기조직화의 최소 의미 이번 프로젝트에서는 아주 좁게 쓴다. SELF-ORGANIZATION 외부 중앙 설계자가 각 구성요소의 배치를 직접 지정하지 않아도, 구성요소의 국소적 상호작용과 동역학으로 거시적 질서·패턴이 형성되는 것. 예:

- 대류패턴
- 결정 성장
- 일부 화학적 패턴
- 군집 패턴

중요: “self”는 의식적 자기라는 뜻이 아니다. 단지: EXTERNAL CENTRAL CONTROLLER ABSENT 라는 뜻에 가깝다. 그래서 첫 번째 HARD LOCK SELF-ORGANIZATION ≠ SELF-AWARE ORGANIZATION 이다. “스스로”라는 한국어 때문에 자꾸 행위자성을 넣게 된다. 하지만 물리적 자기조직화에는:

- 욕망
- 목적
- 자기모델

이 필요 없다. 그러나 이것도 아직 autopoiesis가 아니다 자동온도조절기를 생각하자. 실내온도: 18도. 목표: 22도. 보일러 켜. 상태 회복. 시스템은: 자신의 상태를 조절 한다. 하지만 보일러가:

- 센서
- 배선
- 보일러 부품

을 스스로 생산해 유지하는 것은 아니다. 따라서: **REGULATION $\neq$ AUTOPOIESIS**. 여기서 **autopoiesis**를 원래 의미로 되돌린다 Maturana와 Varela가 제안한 핵심은 매우 특수하다. Autopoietic system은 단순히: 자기를 유지한다. 가 아니다. 핵심은: 구성요소를 생산하는 과정들의 네트워크가 그 네트워크 자체를 계속 재생산하고 동시에 시스템을 하나의 단위로 구성한다 는 것이다. 원래 정의는 상당히 난해하지만, 후속 비판적 리뷰에서도 핵심은 **생산과정의 네트워크가 그 네트워크와 시스템의 경계를 다시 생성한다는 조직적 순환성**으로 정리된다. 이것이 **autopoiesis**의 가장 강한 직관이다 **THE SYSTEM PRODUCES THE CONDITIONS OF ITS OWN CONTINUATION**. 좀 더 정확하게: 자기 자신을 구성하는 생산네트워크와 단위를 계속 재생산한다. 그래서: **SELF-PRODUCTION** 이다. **autopoiesis**는 생명 정의로 성공했는가 여기서는 강하게 말하면 안 된다. Autopoiesis는 생명의 통합적 정의를 제시하려는 야심찬 개념이었다. 그러나 비판적 리뷰는:

- 원래 정의의 모호성
- 실제 적용 어려움
- 생물학 주류에 널리 통합되지 못했다는 점

을 지적한다. 따라서: 는 틀리다. 그래서 **생명에는 최소 세 질문이 있다** Q1 패턴이 스스로 형성되는가? **SELF-ORGANIZATION**. Q2 그 조직을 스스로 유지하는가? **SELF-MAINTENANCE**. Q3 그 조직을 구성하는 생산과 경계까지 계속 재생산하는가? **AUTOPOIESIS** 후보. 이 세 개는 다르다. 이 지도에서 **중요한 것은 AGENCY-0 $\neq$ AGENCY-4** 다. “박테리아도 행위자다.” 라는 문장과 “인간은 책임 있는 행위자다.” 의 agency가 같은 의미가 아니다. 이것이 **biological normativity** 후보 State A $\rightarrow$ 자기유지 가능. State B $\rightarrow$ 자기유지 실패. 그렇다면: **A is better-for-system than B** 라고 기능적으로 말할 수 있다. 이런 자기유지·자율성에서 생물학적 규범성과 agency를 도출하려는 연구프로그램이 존재하지만, 이것이 충분한 규범성 설명인지에는 최근에도 비판과 논쟁이 있다. **앞 장의 identity와 연결된다** 세포는 계속 부품이 바뀐다. 그런데: **경계 안의 조직적 자기생산** 이 지속된다. 따라서: **SELF-PRODUCTION HELPS GENERATE PERSISTENCE**. 즉: **autopoiesis $\rightarrow$ persistence** 라는 강한 연결 후보가 있다. **우주로 가면 더 위험하다** 우주에는: 외부가 있는가? 우주가 자신의 구성요소를 만든다? 별과 은하가 형성된다. 그러면: **universe is autopoietic**. 라고 하고 싶을 수 있다. 하지만 이건 현재: **REJECT**. 왜? Autopoiesis에서 중요한 것은: **특정 조직적 단위가 자신의 경계와 생산네트워크를 재생산하는 것**. 우주 전체에는: **외부 환경과 구별되는 자기생산적 경계** 를 같은 의미로 적용하기 어렵다. 따라서 **HARD LOCK COSMIC SELF-ORGANIZATION $\neq$ COSMIC AUTOPOIESIS**. 별·은하·구조가 자기조직적으로 형성된다고 해서:

우주가 자기 자신을 살아 있는 존재처럼 생산 한다는 뜻은 아니다. 세 종류의 **SELF** **SELF-1 — ORGANIZATIONAL SELF** 자기 경계와 조직을 재생산하는 하나의 시스템. 세포. **SELF-2 — MODELLED SELF** 자신의 상태를 내부적으로 표현. 앞 장의 self-model. **SELF-3 — REFLECTIVE SELF** 자기 목표·가치·정체성을 대상으로 숙고. 앞 장의 reflection. 이 세 층은 완전히 다르다. **A5가 핵심이다** 단순 외부 제어: 센서가 신호 받고 설계자가 정한 명령 수행. 생물: 환경 신호가: **자기 조직의 현재 상태** 와 결합해 의미가 달라진다. 예: 포도당. 배고픈 세포/상태에서는 중요. 충분한 상태에서는 다름. 즉: **SIGNIFICANCE IS SYSTEM-RELATIVE**. 세포가 영양분을 ‘의미 있게’ 취급한다는 표현 약한 의미: 그 신호가 자기유지 반응을 다르게 만든다. 가능. 강한 의미: 세포가 영양분이라는 개념을 이해한다. NO. 따라서: **BIOLOGICAL SIGNIFICANCE $\neq$ CONCEPTUAL UNDERSTAND-**

ING. MASTER MODEL v2.2 기존 질문축: STATE RELATION DYNAMICS CONSTRAINT VALUE MEMORY PERSISTENCE META-REVISION 이제 추가: **AUTONOMY** 시스템이 자신의 존속조건·경계·조직을 어느 정도 내부 과정으로 생성·유지하는가? 그리고 별도: **AGENCY** 환경 변화에 대해 자신의 상태·목표와 연결된 차별적 행동선택을 어느 정도 수행하는가? 두 개를 분리한다. 이것이 처음의 ‘우주의 이치’ 질문에 매우 중요한 방어선이다 우주에서:

- 구조가 생긴다.
- 별이 생긴다.
- 생명이 생긴다.
- 일부 생명이 자기모델을 만든다.
- 일부가 우주를 이해한다.

이 역사에서: **agency appears inside universe**. 그러나: **universe itself therefore has agency** 는 나오지 않는다. **사회도 마찬가지로** 사회가 자신을 재생산한다. 하지만:

- 생물학적 인간
- 자연환경
- 기술
- 외부 사회

에 의존한다. 따라서: **SOCIAL REPRODUCTION ≠ CLOSED SELF-CREATION**. 하지만 ‘선택’도 **의식적 선택이 아니다** 세균이 화학주성을 한다. 농도가 높은 쪽으로 이동한다. 이를: **choice** 라고 기능적으로 부를 수 있지만 인간의 숙고: A와 B 중 왜 A인가? 와 동일하지 않다. 따라서: **BEHAVIORAL SELECTION ≠ DELIBERATIVE CHOICE**. **HARD LOCKS v2.2 자기조직화 ≠ 자기유지 자기유지 ≠ autopoiesis autopoiesis ≠ 재생산 autopoiesis ≠ 생명에 대한 합의된 최종정의 autopoiesis ≠ consciousness autopoiesis ≠ agency 전체 생물학적 autonomy ≠ 환경 독립 self-production ≠ 무로부터 자기창조 functional purpose ≠ conscious intention biological significance ≠ felt value behavioral selection ≠ deliberative choice 사회적 자기재생산 ≠ 세포 autopoiesis AI self-maintenance ≠ 생물학적 autopoiesis 우주 자기조직화 ≠ 우주적 agency** 이 장의 최종 판정 생명의 독특한 점 가운데 하나는 단순히 질서가 있다는 것이 아니라, 그 질서를 가능하게 하는 구성요소와 경계를 자신의 내부 활동을 통해 계속 재생산한다는 데 있다. 그러나 이것은 아직:

- 사고
- 의식
- Reflection
- 도덕성

이 아니다. 그것들은 별도의 추가 구조를 요구한다. 즉: **살아 있음 행위자임 의식함 반성함** 을 더 이상 하나의 단어로 묶지 않는다. 그리고 **아주 중요한 다음 질문이 열린다** 자기유지하는 시스템은: **자기 존속에 유리한 것과 불리한 것** 을 구별한다. 그렇다면: 이 차이가 어떻게 ‘정보’가 되는가? 외부의 화학농도는 그 자체로: **메시지** 가 아니다. 그런데 박테리아에게는: 이쪽으로 가라. 는 행동적 의미를

맞는다. DNA 염기서열도 단순 분자인데: **왜 정보라고 부르는가?** 뉴런 신호도 전압변화인데: **언제 표상이 되는가?**

11장. 정보·표상·모델

물리적 차이는 언제 정보가 되고, 정보는 언제 세계에 관한 의미가 되는가?

차이와 상관은 정보의 출발점일 수 있지만 기능적 중요성, 표상, 의미내용, 경험된 의미는 다른 층위다. 모델은 세계로 가는 통로이지 세계 자체도 감옥도 아니다. 이번 노드는 꽤 중요합니다. 지금까지 우리는 정보라는 말을 너무 자연스럽게 써왔는데, 실제로는 **상관관계가 있다는 것, 신호가 기능적으로 사용된다는 것, 어떤 것을 표상한다는 것, 의미를 가진다는 것**은 서로 다른 주장입니다. 물리적 차이는 언제 '정보'가 되는가 Shannon 정보·신호·생물학적 정보·표상·의미·오류 차이에서 의미로 넘어가는 인터페이스 핵심 사슬은: **DIFFERENCE ↓ CORRELATION ↓ SIGNAL ↓ FUNCTIONAL USE ↓ REPRESENTATION ↓ SEMANTIC CONTENT** 이다. 그러나 이 역시 자동적인 계단이 아니다. 각 화살표마다 별도의 인터페이스가 필요하다. 2026년 최신 생물학적 정보 개관도 바로 이 점을 강조한다. Shannon 정보는 비교적 명확하고 넓게 적용되지만, 생물학에서 흔히 말하는 '의미 있는 정보'는 비대칭성·정확성과 오류 가능성·기능 같은 더 강한 조건을 요구하며 여전히 논쟁적이다. **첫 번째 사례 나무의 나이테** 나무를 자른다. 나이테가 많다. 우리는: 이 나무는 오래됐다. 라고 추론한다. 즉: tree rings 와 tree age 사이에 상관관계가 있다. 따라서 나이테는: **나무 나이에 관한 정보를 가진다** 라고 말할 수 있다. 그런데 나무가 그 정보를 '읽는가' 대개 아니다. 나무는 자신의 나이테를 보고: 내가 73살이군. 이라고 판단하지 않는다. 나이테는: **우리에게 정보** 를 제공한다. 하지만: **나무에게 사용되는 정보** 라는 것은 별도 문제다. 이 차이는 철학적 생물학 정보 논의의 고전적 예다. Shannon적 의미에서는 나이테와 나이 사이의 상관만으로 정보라고 부를 수 있지만, 그 정보가 나무의 내부 제어에 사용된다는 뜻은 아니다. **Shannon information은 아주 강력하지만 매우 얇다**

Claude Shannon의 정보이론에서 핵심은: **불확실성** 이다. 어떤 변수 X가 여러 가능한 상태를 가진다. X를 관측하기 전: 불확실하다. 신호 Y를 알면: X에 대한 불확실성이 줄어든다. 그러면: **Y carries information about X.** 라고 말할 수 있다. 그런데 이 수식에는 '뜻'이 없다 매우 중요하다. 010011010... 이라는 비트열이 있다. Shannon 이론은:

- 얼마나 압축 가능한가
- 얼마나 불확실성을 줄이는가
- 채널이 얼마나 많은 정보를 전송하는가

를 다룰 수 있다. 하지만: 이 비트열이 “내일 비가 온다”는 뜻인가? 는 별도 문제다. 즉: **SHANNON INFORMATION ≠ SEMANTIC MEANING.** 이다. SEP의 생물학적 정보 정리도 Shannon 정보가 상관·불확실성에 관한 최소적 개념이며, 생물학에서 사용하는 더 풍부한 정보 개념과는 구별된다고 정리한다. **이건 매우 중요한 HARD LOCK MORE INFORMATION ≠ MORE MEANING.** 랜덤 문자열은: Shannon entropy가 높을 수 있다. 하지만: **아무 의미 없는 잡음** 일 수도 있다. 반대로: “불” 두 글자는 매우 짧다. 하지만 상황에 따라: **엄청난 의미** 를 가진다. **두 번째 사례 연기** 멀리 연기가 보인다. 연기: **불과 상관.** 따라서: 불에 관한 정보. 그런데 연기가: 불을 의미하려고 생긴 것은 아니다. 따라서:

NATURAL CORRELATION 이다. 앞 장와 바로 연결된다 박테리아가 화학물질 농도를 감지한다. 수용체. 신호전달. 편모운동 변화. 즉: **environmental difference** → **biological response**. 여기서 화학농도는: **박테리아의 행동을 바꾸는 신호** 다. 세 번째 사례 DNA DNA는 왜 '정보'를 가진다고 말할까? 문자열: A T G C ... 염기서열. 이 서열이:

- RNA로 전사되고
- 단백질 번역과 연결된다.

특히 유전암호에서는: **nucleotide triplet** ↔ **amino acid** 대응규칙이 있다. 그래서: **CODE** 라는 강한 정보언어가 생겼다. 그렇다고 유전정보가 완전히 비유인가 그것도 지나치다. 특히: **genetic code** 에는 매우 특수한 매핑구조가 있다. 염기서열 변화가: 아미노산 서열과 체계적으로 대응한다. 따라서 이 영역에서는: **정보·코드 언어** 가 단순 시적 비유보다 훨씬 강한 근거를 가진다. **생물학적 정보의 큰 철학적 질문** 어떤 상태가 다른 상태와 상관되기만 하면: **정보** 라고 부를 수 있다. 그런데 생물학자들이 실제로 원하는 것은 더 강하다. 예: 경보음이: 포식자가 있다. 를 나타낸다. 그렇다면 경보는: **틀릴 수도 있어야 한다**. 포식자가 없는데 경보. **FALSE ALARM**. 여기서: **ERROR** 가 등장한다. 왜 이게 중요한가 온도계: 20°C 표시. 실제: 25°C. 우리는: 온도계가 잘못 표시했다. 고 말한다. 즉 표시값에는: **correctness condition** 이 있다. 그런 점에서는 단순 상관보다 표상적 성격이 강해진다. **표상(representation)**은 과학에서도 하나로 정의되지 않는다 이 점도 아주 중요하다. 2025년 *Mind & Language* 논의를 뇌과학에서 representation이라는 개념이 널리 사용되지만 정확한 의미가 통일되어 있지 않다고 지적한다. 2026년 *Nature Reviews Neuroscience*는 이 혼란을 줄이기 위해 신경표상을 몇 가지 차원으로 나누자고 제안한다. 예:

- 특정 feature에 민감한가
- 그 feature에 특이적인가
- 다른 변화에도 불변성을 갖는가
- downstream system이 그 신호를 실제로 사용하는가.

즉: **DECODABLE** ≠ **REPRESENTED** 라는 문제가 실제 신경과학에서도 존재한다. 강한 **HARD LOCK DECODABILITY** ≠ **FUNCTIONAL REPRESENTATION**. 이다. 어떤 변수에서 정보를 통계적으로 읽어낼 수 있다고: **그 시스템이 그 정보를 사용한다** 는 뜻은 아니다. 이것은 AI interpretability에서도 중요하다. 이 판정표로 연기를 검사하자 R1: 불과 상관 있음. R2: 부분적으로. R3: 상황 따라 약함. R4: 어떤 시스템이 사용할 수는 있음. R5: 연기 자체에는 자연적으로 '거짓'이라는 규범이 없음. 따라서: **INFORMATION SOURCE** 는 강하지만 **REPRESENTATION** 은 약하다. **AI를 넣어보자** 언어모델에: “파리는 프랑스의 수도다.” 라는 입력. 내부 활성 변화. 출력: “맞습니다.” 기능적으로:

- semantic relationships
- world-related regularities

를 처리한다. 따라서: **REPRESENTATIONAL FUNCTION** 이라는 말은 충분히 검토할 가치가 있다. 그리고 '이해'라는 말을 쓰려면 더 조심해야 한다 AI가 올바르게 답한다. 그렇다면: 이해했다. 고 말할 수 있는가? '이해'도 여러 의미가 있다.

- 관계를 적절히 사용
- 새로운 상황에 일반화
- 인과적으로 설명
- 경험적으로 의미를 느낌

전부 다르다. 따라서: **CORRECT OUTPUT $\neq$ SINGLE SETTLED CONCEPT OF UNDERSTANDING**. 이다. 그러면 **KNOWLEDGE**는? 철학에서 지식 자체가 거대한 주제지만 이번 프로젝트에서는 최소한: **INFORMATION $\neq$ KNOWLEDGE** 만 유지한다. 정보를 받았다고:

- 믿었다
- 이해했다
- 정당하게 믿었다
- 실제 행동에 사용할 수 있다

가 자동으로 나오지 않는다. 과학은 특히 이 변환을 제도화한다 실험결과: **data**. 그 결과가 기존 이론과 다름: **signal**. 통계·방법론 검증: **evidence**. 이론경쟁: **interpretation**. 재현·동료검토: **knowledge filtering**. 즉 과학의 핵심은: 많은 정보를 생산 하는 것보다 어떤 차이를 믿을 만한 근거로 승격시킬지 판별 하는 데 있다. 즉 정보가 풍부할수록 새로운 병목이 생긴다 과거: **INFORMATION SCARCITY**. 미래: **ATTENTION / VALIDATION / TRUST SCARCITY**. 일 수 있다. 이것은 현재 프로젝트의 강한 사회적 함의다. 정보가 많다고 지능이 높은 것이 아니다 가능한 수를: 백만 개 생성. 하지만 좋은 수를 구별 못함. 지능 낮음. 반대로: 핵심 후보 3개만 보고도 좋은 선택. 전문가. 따라서: **INFORMATION ACCESS $\neq$ INTELLIGENCE**. 왜냐하면 **task**가 없으면? 인간은: 예술을 본다. 목적 없는 관찰. 상상. 꿈. 표상이 꼭 즉각적 task에만 사용되는 것은 아니다. 따라서: **TASK-RELEVANCE $\neq$ FULL REPRESENTATION THEORY**. **ERROR FAMILY**를 다시 분리한다 *E1 — CONTROL ERROR* 목표상태 vs 현재상태. *E2 — PREDICTION ERROR* 예측 vs 관측. *E3 — REPRESENTATIONAL ERROR* 표상 내용 vs 세계. *E4 — NORMATIVE ERROR* 행동·규칙 vs 가치기준. 앞 장의 오류분류가 더 정교해진다. 이것이 범용 AI 논의에서도 중요하다 AI가 자기출력 오류를 감지한다. 질문: 어떤 오류?

- 확률적으로 예상과 다름?
- 외부 사실과 불일치?
- 사용자의 목표에 안 맞음?
- 도덕적으로 문제?

서로 다른 평가채널이다. 따라서: **SELF-CORRECTION** 이라는 말도 반드시 기준을 명시해야 한다. 이것은 뇌과학에서 특히 중요하다 뉴런 A 활동과 행동 B 상관. decode 가능. 그런데: A를 조작해도 B 변화 없음. 그러면: **causal relevance** 가 약할 수 있다. 따라서 강한 설명에는: **CORRELATION + INTERVENTION** 이 중요해진다. 이제 이 프로젝트에서 '정보'를 사용할 때 규칙을 잠근다 *INFORMATION-1 SHANNON / CORRELATIONAL* X를 알면 Y 불확실성 감소. *INFORMATION-2 SIGNAL / FUNCTIONAL* X가 시스템 내부 전달경로에서 사용됨. *INFORMATION-3 REPRESENTATIONAL* X가 Y를 나타내며 기능적으로 사용되고 오류 가능. *INFORMATION-4 SEMANTIC* 명시적 내용·참조조건·지시대상 등을 가짐. *INFORMATION-5 AGENT SIGNIFICANCE* 그 내용이 행위자의

가치·행동·경험에 어떤 차이를 만드는가. 앞으로 정보라는 말을 쓰면: 어느 층인지 암묵적으로라도 확인한다. 이것은 앞 장 시장논의에 중요하다 시장은: **signal generator** 로는 매우 강하다. 가격이 행동을 바꾼다. 하지만: **perfect truth representation** 은 아니다. 즉: **MARKET SIGNAL ≠ TRUE SOCIAL VALUE**. 이전 HARD LOCK들이 더 엄격해진다. 법원 판결도 판결: 법적 판단 신호. 하지만: 진실 그 자체. 는 아니다. 항소 가능. 즉 제도는: **imperfect representations** 을 다루는 교정체계가기도 하다. 매우 중요한 후보명제 지능의 핵심은 정보를 많이 갖는 것이 아니라, 어떤 차이가 현재 목표·모델·가치에 의미 있는 차이인지 구별하는 능력일 수 있다. 이 문장은 꽤 강하다. 하지만: 지능의 완전한 정의 라고 잠그지는 않는다.

그리고 AI는 **representation**을 학습한다 raw board: 그런데 내부적으로: **future-relevant structure** 를 추출한다. 즉: **data compression into action-relevant latent structure**. 이것이 머신러닝 representation learning의 핵심 직관과 연결된다. 이것은 앞 장의 **LEARNING** 정의를 확장한다 기존: **past interaction → future policy change**. 이제: **past interaction → representation change + policy change**. 즉: **LEARNING MAY CHANGE WHAT THE SYSTEM SEES AS RELEVANT**. 이것이 매우 중요하다. 따라서 정보는 절대 독립적이지 않다는 말을 조심스럽게 이해할 수 있다 이 문장은 흔히 남용된다. 물리적 데이터 자체의 진위를: 다 가치판단이다. 라고 할 수는 없다. 하지만: 어떤 데이터를 수집하고, 어떤 변수를 모델에 넣고, 무엇을 중요한 **signal**로 보는가 에는 목표와 제도가 영향을 줄 수 있다. 즉: **DATA ≠ VALUE-FREE SELECTION ENVIRONMENT**. 라는 좁은 의미는 강하다. **MASTER MODEL v2.3 — INFORMATION AXIS** 이제 기존 다섯 축에 하나를 명확히 추가한다. **AXIS F — INFORMATION / REPRESENTATION F0 Difference** 상태차이. *F1 Correlation* 통계적 정보. *F2 Signal* 전달·기능적 사용. *F3 Significance* 시스템 존속·목표에 대한 중요성. *F4 Representation* 무엇인가를 나타내며 오류 가능. *F5 Semantic Content* 무엇에 관한 것인지 구조화된 내용. *F6 Experienced Meaning* 행위자에게 경험적으로 어떤 의미인가. 이것은 '정보가 우주의 근본이다'라는 주장도 막아준다

현대 물리학에서 정보 개념은 매우 중요하다. 양자정보. 블랙홀 정보. 열역학. 그렇다고: **universe is literally made of information** 이라는 존재론이 자동으로 나오지 않는다. 왜냐하면: **Shannon information itself presupposes possible states and probability structure**. 정보가: 무엇의 정보인가 가 여전히 남는다. 따라서 **HARD LOCK INFORMATIONAL DESCRIPTION ≠ INFORMATIONAL ONTOLOGY**. 이다. 정보로 세계를 훌륭하게 기술할 수 있다는 것과: 세계의 물질적 실재가 정보뿐이다 는 것은 다른 주장이다. 앞 장의 structural realism과 같은 경계다. 그리고 이것은 AI에도 동일하다 AI는 training data에서 세계를 배운다. 인간은: 감각과 경험에서 세계를 배운다. 둘 다: **MODEL ≠ WORLD**. 그러면: 어떤 모델이 더 참에 가까운가? 를 판정해야 한다. 여기에서: **prediction intervention error correction scientific realism** 이 들어온다. 이 장의 **HARD LOCKS** 차이 ≠ 정보 Shannon 정보 ≠ 의미 정보량 ≠ 중요성 상관 ≠ 신호 신호 ≠ 표상 표상 ≠ 이해 decode 가능 ≠ 시스템이 실제 사용 representation ≠ consciousness DNA 정보 ≠ 완전한 생물 청사진 정보 ≠ 지식 정보 ≠ 가치 정보관계 ≠ 인과관계 representation error ≠ prediction error ≠ normative error 정보적 기술 ≠ 정보가 존재론적으로 근본 이 장의 가장 강한 결론

세상에는 수없이 많은 차이와 상관관계가 있지만, 어떤 차이가 시스템 내부에서 신호로 사용되고, 무엇인가를 나타내며, 틀릴 수 있고, 행동과 가치에 차이를 만들기 시작할 때 우리는 점점 더 강한 의미의 '정보'를 말할 수 있다. 따라서: **INFORMATION** 은 한 단어가 아니다. 적어도: 통계적 정보

기능적 신호 표상적 내용 행위자에게 중요한 의미를 분리해야 한다. 그리고 바둑에서 시작한 질문이 다시 돌아온다 초보자와 신진서는: 같은 돌을 본다. 하지만 같은 정보를 갖는다고 말하기 어렵다. 더 정확하게는: 같은 물리적 입력에서 서로 다른 구조를 표상하고 서로 다른 미래관련 정보를 추출한다. 즉: 전문성이란 정보량 증가보다 어떤 차이를 의미 있는 차이로 압축하는 능력일 수 있다. 이번 노트에서는 “잘 예측한다”와 “세계를 안다”를 분리합니다. 과학철학에서도 모델은 현실의 복사본이라기보다 특정 목적을 위해 현실의 일부 구조를 골라 표현하는 도구로 다뤄집니다. 2025년 개정된 과학모델 문헌도 좋은 모델이 현실을 그대로 닮아야 하는 것은 아니며, 이상화된 모델도 실제 대상에 관한 지식을 줄 수 있다고 정리합니다. 우리는 세계를 보는가, 세계에 대한 모델을 보는가 예측 · 표상 · Bayesian inference · 실험 · 인과 · 과학적 실재론 모델과 현실 사이에는 무엇이 있는가 가장 중요한 구분은: MODEL-MEDIATED ACCESS ≠ MODEL-IMPRISONMENT 이다. 우리는 세계를: 모델 없이 보지 않을 수 있다. 그러나 모델은: 세계와의 접촉을 통해 계속 실패할 수 있다. 바로 그 실패 가능성이 중요하다. 다시 바둑 AI부터 시작하자 특정 엔진·모델·규칙·덤·탐색 설정을 고정한 분석 화면에 다음 값이 표시된다. 흑 승률 63%. 질문: 63%는 바둑판 어디에 존재하는가? 돌 위? 아니다. 빈 공간? 아니다. 63%는: 현재 판 상태를 모델에 넣었을 때 나온 평가값 이다. 즉:

WORLD STATE 와 MODEL ESTIMATE 를 분리해야 한다. 바둑판은 현실이다 현재 돌 배치: 실제로 존재한다. 그런데: 이 판은 흑이 63% 이긴다. 는 것은:

- 어떤 플레이어 수준
- 어떤 탐색
- 어떤 정책
- 어떤 평가모델

을 전제로 한다. 따라서: POSITION ≠ WIN PROBABILITY. 승률은 조건부 주장이다 더 정확하게: $P(\text{Win} \mid \text{State, Model, Policy assumptions})$ 에 가깝다. 즉: 세계의 절대적 숫자 가 아니라 특정 모델 아래의 조건부 예측 이다. 그렇다면 모델은 가짜인가 아니다. 지도도: 영토가 아니다. 그러나 좋은 지도는:

- 길을 찾게 하고
- 거리를 추정하게 하고
- 목적지에 도달하게 한다.

과학모델도 이와 비슷한 점이 있다. 다만 과학철학에서 모델은 지도보다 훨씬 다양하며, 현실의 일부를 의도적으로 왜곡하거나 이상화해도 설명·예측에 유용할 수 있다. 첫 번째 HARD LOCK MODEL ≠ WORLD. 하지만 동시에: MODEL ≠ ARBITRARY FICTION. 이다. 좋은 모델은: 세계의 어떤 구조와 체계적으로 맞물려야 한다. 그렇지 않으면 실패한다. 예측은 모델을 시험하는 첫 번째 인터페이스다 모델: 내일 비. 현실: 맑음. 그러면: prediction error. 발생. 즉 모델은 자신의 내부 논리만으로 살아남지 않는다. WORLD CAN SAY NO. 이것이 매우 중요하다. 예측오류가 왜 그렇게 중요한가 모델이 항상: 내가 틀린 것이 아니라 상황이 특별했다. 라고 말할 수 있다면? 어떤 결과도 반증이 되지 않는다. 그러면: 학습 불가능. 따라서 좋은 모델은 어느 정도: 위험을 감수해야 한다. 즉: 이 조건에서는 이런 결과가 나올 것이다. 라는 제한을 둔다. 그런데 예측만 잘하면 참인가 여기서 바로 문제가 생긴다. 두 모델 M1, M2가 있다. 둘 다: 동일한 관측결과 를 예측한다. 그런데

내부 설명은 다르다. 그러면: 어느 쪽이 진짜 세계구조인가? 를 예측만으로 결정하기 어려울 수 있다. **예측성공만으로 존재론을 확정할 수 없는 이유** 모델 A: 보이지 않는 입자 X가 있다. 모델 B: 입자 없이 다른 수학구조로 설명. 둘 다 같은 실험결과. 그러면: **prediction success** 만으로: X는 실제로 존재한다. 를 확정하기 어렵다. 즉: **PREDICTIVE ADEQUACY $\neq$ ONTOLOGICAL TRUTH**. 이 프로젝트에서는 어느 쪽을 잠그는가 둘 중 하나를 최종 확정하지 않는다. 대신: **EMPIRICAL SUCCESS** 와 **REALIST INTERPRETATION** 을 분리한다. 이것만으로도 충분히 중요하다. 왜 개입이 강력한가 우산 판매와 비가 상관된다.

우산 판매를 강제로 늘렸다고: 비가 오는가? 아니다. 따라서: **correlation** 만으로 원인을 알 수 없다. 따라서 **MODEL-WORLD CONTACT**에는 최소 세 단계가 있다 **OBSERVATION** 세계를 본다. **PREDICTION** 모델로 아직 보지 않은 결과를 예상한다. **INTERVENTION** 세계를 바꾸고 결과를 본다. 셋은 증거력이 다르다. 이것이 **agency**와 연결된다 행위자는: 세계를 관찰만 하지 않는다. 행동한다. 즉: Model $\downarrow$ Action $\downarrow$ World changes $\downarrow$ Observation $\downarrow$ Model update 이다. 앞 장의 페루프가 다시 나타난다. 바둑에서도 같다 프로가 수를 둔다. 그 수는: **승부행동** 이면서 동시에 **상대 반응에 대한 실험** 이 될 수 있다. 저쪽이 어떻게 받는지 보자. 즉: **ACTION CAN HAVE INFORMATION VALUE**. 앞 장의 핵심이 또 돌아온다. **인간 학습도 이런 틀로 모델링할 수 있다** 최근 인간 학습 연구는 불확실성의 종류에 따라 prediction error를 얼마나 크게 반영할지를 조정하는 Bayesian predictive inference 틀을 활발히 사용한다. 그리고 뇌의 prediction-error 회로가 불확실성까지 추정할 수 있다는 2025년 모델도 제시됐다. 오히려 **prediction error**가 외부세계의 독립성을 보여주는 중요한 인터페이스다 내 모델: 문이 열려 있다. 걷는다. 쿵. 실제: 닫혀 있다. 세계는: **내 믿음에 맞춰주지 않는다**. 즉: **RESISTANCE OF WORLD** 가 있다. 이것이 강한 실재론적 직관을 만든다 세계가 내 모델과 독립적이지 않다면: 왜 계속 놀라는가? 왜 예측이 틀리는가? 왜 실험결과가 마음대로 나오지 않는가? 즉: **ERROR IS EVIDENCE OF MODEL-WORLD NONIDENTITY**. 이다. 하지만 **prediction error**만으로 어느 모델이 참인지는 해결되지 않는다

모델이 틀렸다. 가능성:

- 기본가설 틀림
- 보조가설 틀림
- 측정 오류
- 실험 설계 오류

즉: **ERROR SIGNAL $\neq$ CAUSE OF ERROR**. 앞 장의 원칙이 또 등장한다. AI도 똑같다 AI에게 후보진단 5개를 줬다. 아무리 정교하게 확률을 계산해도: **실제 병이 목록 밖** 이면 실패한다. 따라서: **UNCERTAINTY OVER KNOWN OPTIONS** 와 **UNKNOWN OPTION** 은 다르다. 바둑도 완전히 예외는 아니다 합법수는 모두 정의돼 있다. 하지만 인간 기사에게: **실제로 고려 가능한 패턴** 은 제한적이다. “생각하지 못한 수.” 가 존재한다. 즉 법적 option space는 닫혀 있어도: **cognitive hypothesis space** 는 부분적으로 열려 있다. 이세돌 78수의 충격도 이런 관점에서 이해할 수 있다. 모델은 오히려 일부러 틀릴 수 있다 이상기체. 마찰 없는 표면. 완전히 합리적인 경제인. 현실에 그대로 존재하지 않는다. 하지만: 특정 구조를 고립해서 이해 하는 데 유용하다. 즉: **LITERAL FALSEHOOD CAN COEXIST WITH EXPLANATORY USEFULNESS**. 과학모델 문헌이 이상화와 현실적 표현을 엄격히 동일시하지 않는 이유다. 매우 중요한 **HARD LOCK MORE REALISTIC MODEL $\neq$ AL-**

WAYS BETTER MODEL. 이다. 지도에:

- 모든 나무
- 모든 창문
- 모든 사람

을 넣으면? 정확도는 늘어날 수 있지만: **길찾기 기능은 나빠질 수 있다.** 좋은 모델은: **필요한 차이를 보존하고 불필요한 차이를 버린다.** 앞 장의 **compression**이 또 돌아온다 좋은 모델: **reality compression.** 하지만: **relevant structure preserved.** 즉: 이라는 강한 방법론적 직관이 생긴다. 따라서 모델의 '정확성'도 단일 숫자가 아니다 예: 기후모델. 어떤 모델은: 평균기온 예측 좋음. 지역강수 예측 약함. 또 다른 모델은 반대. 즉: **ACCURATE WITH RESPECT TO WHAT?** 를 물어야 한다. 외부검증의 철학적 의미 훈련데이터: **model's familiar world.** 새 병원: **independent reality test.** 모델이 진짜 구조를 잡았다면: **transfer.** 표면적 상관만 외었다면: **collapse.** 따라서: **GENERALIZATION IS A MODEL-WORLD TEST.** 이것이 회의주의와 실재론 사이의 중요한 중간지점 우리는 완전한 확실성을 갖지 못할 수 있다. 하지만: **아무것도 모르는 것도 아니다.** 모델끼리:

- 예측
- 실험
- 외부검증
- 재현

으로 성능차이를 만들 수 있다. 즉: **FALLIBILITY ≠ EPISTEMIC NIHILISM.** 따라서 모델에는 세계가 부과하는 제약이 있다 좋은 표현: **REALITY CONSTRAINS MODEL SPACE.** 모든 이야기가 살아남는 것은 아니다. 세계가: **일부 모델을 죽인다.** 여기서 “한 가지 진실”과 “여러 모델”이 모순인가 반드시 아니다. 하나의 산을: 북쪽 지도. 지질도. 등고선도. 모두 다르게 그린다. 대상: **하나.** 표상: **여러 개.** 따라서: **MULTIPLE REPRESENTATIONS ≠ MULTIPLE REALITIES.** 이다. 이제 인간 지각으로 돌아가자 눈에 빛이 들어온다. 뇌가 처리. 우리가: **세계를 경험.** predictive-processing 계열에서는 뇌가 noisy sensory input과 prior expectations를 결합하며 prediction error를 이용해 추론을 조정한다고 본다. 그렇다면: **perception is constructive.** 착시가 왜 중요한가 착시는: **perception ≠ raw world** 를 보여준다. 그러나 동시에: 평소에는 이 추론체계가 꽤 잘 작동 한다는 사실도 보여준다. 착시가 놀라운 이유는: **정상적인 추론이 대부분 안정적 이기 때문이다.** 인간은 '현실'을 보는가 가장 안전한 답: **우리는 현실에 의해 인과적으로 제약된, 내부적으로 구성된 경험을 가진다.** 즉: **DIRECT COPY** 도 아니고 **FREE FANTASY** 도 아니다. AI agent가 행동하면 달라진다 로봇. 웹 에이전트. 실험자동화 AI. 행동: **intervention.** 결과: **feedback.** 그러면 AI도: **model → action → world → observation** 폐루프에 들어간다. 즉 단순 language prediction보다 강한 world-model testing이 가능해진다. 이제 과학이 왜 특별한지 더 정확히 볼 수 있다

개인의 지각도 모델을 수정한다. 그런데 과학은 여기에:

- 공개측정
- 재현
- 경쟁가설
- 독립검증

- 도구교정

을 붙인다. 즉: PERSONAL MODEL CORRECTION 을 INSTITUTIONAL MODEL CORRECTION 으로 확장한다. 하지만 과학도 현실의 최종 거울은 아니다 모델은:

- 이상화
- 제한된 관측
- 역사적 기술

을 갖는다. 그래서 과학은: finished map 보다 corrigible mapping process 로 보는 편이 더 강하다. 프로젝트 판정 STRONG REALISM “현재 과학모델이 세계 그대로다.” REJECT. STRONG SKEPTICISM “모델뿐이므로 세계에 대해 아무것도 알 수 없다.” REJECT. CRITICAL / FALLIBILIST REALISM CANDIDATE 모델은 현실과 동일하지 않지만 현실에 의해 제약되며, 반복적인 예측·개입·검증을 통해 세계구조의 일부를 점점 더 안정적으로 포착할 수 있다. RETAIN AS PROJECT STANCE. 철학적 학파를 최종 선택했다는 뜻은 아니다. 이것은 인생에도 매우 중요한 태도다 나는: 나에 대한 모델 을 가진다. 나는 숫자에 약해. 나는 사람을 잘 본다. 이 관계는 안 될 거야. 이진: self-model. 현실 그 자체가 아니다. 그런데 자기모델은 더 복잡하다 내가: 나는 이런 사람이다. 라고 믿으면 그 믿음: 내 행동을 바꾸고 결국 그 사람이 될 수도 있다. 사회적 reflexivity가 자기모델에도 나타난다. 그래서 인간 예측에는 반사성이 있다 예측: 행동변화. 행동변화: 예측 대상 변화. 즉: PREDICTION CAN BECOME AN INTERVENTION. 금융시장. 정치여론조사. 개인상담. 등에서 매우 중요하다. 질문 하나가 추가된다 WHAT DOES THE PREDICTION DO TO THE WORLD? 이다. 즉 예측모델은: passive observer 가 아니라 causal participant 가 될 수 있다. 그리고 세 번째가 있다 REFLEXIVE ACCURACY 내 예측을 공개하거나 행동에 사용했을 때: 대상 시스템이 어떻게 반응하는가? 사회에서는 필요하다. 바둑에는 상대가 내 전략을 읽는 문제 정도에서 부분적으로 나타난다. 이제 승률을 다시 보자

같은 설정에서 KataGo가 표시한 값: 63%. 이 숫자를 본 인간 기사: 심리 변화. 플레이 변화. 그러면: 공개된 승률 자체가 게임 결과에 영향을 줄 수 있다. AI 평가가: model output 에서 environment input 이 된다. 흥미로운 페루프다. 하지만 이 사실 때문에 객관적 사실이 사라지는 것은 아니다 다시 HARD LOCK: REFLEXIVITY ≠ EVERYTHING IS SUBJECTIVE. 모델이 현실에 영향을 줄 수 있다는 것과: 모든 사실이 모델에 의해 만들어진다는 것은 다르다. 모델이 대상에 영향을 주는 시스템에서는 한 화살표 추가: MODEL OUTPUT → TARGET BEHAVIOR 즉: REFLEXIVE LOOP. 이다. 이 장의 HARD LOCKS 모델 ≠ 세계 모델을 통한 접근 ≠ 모델 안의 감옥 예측 성공 ≠ 존재론적 진실 경험적 적합성 ≠ 유일한 참이론 동일한 예측 ≠ 동일한 설명 상관 ≠ 인과 관찰 ≠ 개입 Bayesian update ≠ 새 가설 생성 높은 사전확률 ≠ 진리 현실적 모델 ≠ 항상 좋은 모델 이상화 ≠ 쓸모없는 거짓 일반화 성공 ≠ 절대적 진리 constructive perception ≠ 외부세계 부정 세계모델 ≠ 의식 예측 ≠ 독립적 관찰, 특히 사회에서는 이 장의 가장 강한 결론 우리는 세계와 모델을 동일하게 가질 수는 없지만, 세계가 모델에 저항하고 예측을 실패시키며 개입에 다르게 반응하기 때문에 모델을 서로 비교하고 수정할 수 있다. 따라서 우리가 세계를 안다는 것은: 세계의 완벽한 복사본을 머리에 넣는 것이 아니라 세계와 반복적으로 부딪혀도 살아남는 표상과 모델을 만들어가는 과정 에 더 가깝다. 그리고 ‘진리에 가까워진다’는 말도 새롭게 보인다 단순히: 예측 적중률 ↑ 만은 아니다. 더 강한 후보:

- 더 넓은 환경에서 성공
- 새로운 예측 성공
- 실제 개입결과 설명
- 서로 다른 현상을 통합
- 오류 발생 시 수정 가능

이다. 즉: **ROBUST CONTACT WITH REALITY**. 이것이 진리 그 자체의 정의는 아니지만: **좋은 인식모델의 강한 지표** 다. 인생에서는 한 단계 더 어렵다 삶에서는:

- 목표가 바뀌고
- 행동이 자신을 바꾸고
- 예측이 행동을 바꾸며
- 환경도 반응한다.

따라서 인생모델에는: **prediction + intervention + self-transformation + reflexivity** 가 모두 들어간다. 그래서: **LIFE IS NOT MERELY A HARDER PREDICTION PROBLEM**. 이다. 그리고 여기서 가장 깊은 문제가 열린다 우리는 모델을 수정한다. 과학도 수정한다. AI도 수정한다. 그런데 왜: 어떤 모델을 더 '참'이라고 해야 하는가? 단순 예측보다 더 깊게 가면: **TRUTH** 자체가 등장한다. “참”은:

- 대응인가?
- 일관성인가?
- 실용적 성공인가?
- 과학적 실재와 어떤 관계인가?

12장. 참·현실접촉·의미

진실이 항상 편안하지 않은데도 우리는 왜 진실과 실제 세계의 관계를 가치 있게 여기는가?

참, 최적, 옳음, 정당성은 서로 다른 평가다. 진실은 자율성, 실제 관계, 실제 성취와 이해를 가능하게 하며, 의미는 좋은 느낌만으로 환원되지 않는다. '참'이라는 것은 정확히 무엇인가 대응·일관성·실용적 성공·예측·과학적 실재 *FACT*·*OPTIMUM*·*OUGHT*는 같은 종류의 '맞음'인가 이 장의 핵심은 세 종류의 "맞다"를 분리하는 것이다. **TRUE** 세계가 실제로 어떠한가에 대한 판단. **OPTIMAL** 주어진 목표와 제약 아래 어떤 선택이 가장 나은가에 대한 판단. **RIGHT** 어떤 행동·규칙이 정당화될 수 있는가에 대한 규범적 판단. 이 셋을 한 축으로 놓으면 프로젝트 전체가 다시 흐려진다. 세 문장을 보자 문장 A 물은 표준대기압에서 약 100°C에서 끓는다. 문장 B 이 바둑판에서는 A수가 가장 좋은 수다. 문장 C 약자를 착취해서는 안 된다. 우리는 세 문장 모두에: 맞다. 라고 말할 수 있다. 그런데 정말 같은 종류의 맞음인가? 아니다. A는 사실판단이다 질문: 세계가 실제로 그런가? 실험할 수 있다. 조건을 정한다. 측정한다. 결과를 비교한다. 즉: **CLAIM** ↔ **WORLD** 관계가 중요하다. 이것이 진리의 가장 고전적인 직관이다. 하지만 '대응'은 생각보다 어렵다 문장: 전자는 스핀 1/2을 갖는다. 무엇과 어떻게 대응하는가? 눈앞에: 스핀 1/2 물체가 보이는 것은 아니다. 과학이론은:

- 수학
- 추상구조
- 보이지 않는 존재

를 포함한다. 따라서: **correspondence** 의 정확한 메커니즘을 설명하는 것은 쉽지 않다. 하지만 **coherence**만으로 **truth**가 될까 두 개의 완전히 다른 소설 세계가 있다. 각각 내부적으로 완벽하게 일관적. 그래도 둘이: 현실에 관해 모두 참 일 필요는 없다. 따라서: **COHERENCE** ≠ **CORRESPONDENCE**. 이다. 하지만 유용함: 참? 아니다. 거짓 믿음도 유용할 수 있다. 예: 복잡한 현실을 무시한 단순모델. 정확히 참은 아니지만: 특정 문제 해결에는 훌륭하다. 뉴턴역학이 극단적 조건에서는 정확하지 않아도 많은 공학문제에 유용한 것과 비슷하다. 따라서: **USEFUL** ≠ **TRUE**. 즉 진리철학 자체도 합의가 끝난 분야가 아니다 후보:

- 대응
- 일관
- 실용
- deflationary approaches

등이 존재한다. 따라서 이 프로젝트에서는 진리의 단일한 최종 형이상학적 정의를 잠그지 않는다. 최적성에는 목표가 필요하다 최적성은 목표와 제약 아래 어떤 행동이 가장 나은지를 묻는다. 따라서 목적함수 없이는 최적수도 없다. 바둑에서는 목표가 거의 고정되어 있다 WIN. 그래서: 좋은 수. 가 비교적 명확하다. 하지만 삶에서는:

- 돈
- 건강
- 시간
- 관계
- 의미

가 충돌한다. 그래서 앞 장에서: 목적함수 자체가 문제 가 됐다. 중요한 **HARD LOCK OPTIMAL $\neq$ TRUE**. “A수가 최선이다.” 는: **사실적 요소를 가진 평가판단** 이지만 “물은 H2O다.” 와 같은 종류의 명제는 아니다. 최적성은: **목표에 의존한다**. 하지만 **objective-relative: subjective는 아니다** 목표가: 최단시간. 으로 주어졌다면 경로 길이와 교통상태는: **객관적으로 비교 가능** 하다. 즉: **GOAL-DEPENDENT $\neq$ ARBITRARY**. 최적화도 사실을 필요로 한다 목표: 서울에서 부산까지 가장 빠르게 이동. 세계모델이 틀리면: **잘못된 경로**. 따라서: **OPTIMIZATION REQUIRES TRUE-ENOUGH MODEL**. 즉 truth는: **optimization input**. 하지만 optimum 자체는 truth와 같지 않다. 예 — **의료 AI** AI: 치료 A가 생존율을 5% 높입니다. **EPISTEMIC CLAIM**. 그렇다면: 생존을 최대화하려면 A를 선택. **INSTRUMENTAL CLAIM**. 하지만:

- 부작용
- 환자 선택
- 삶의 질
- 비용

을 고려하면: A를 시행해야 하는가? **NORMATIVE / CLINICAL DECISION**. 세 층을 섞으면 안 된다. 이 구조가 **alignment 문제의 핵심**을 다시 보여준다 AI가 완벽한 world model을 가진다고 하자. **EPISTEMIC CAPACITY MAX**. 그래도: **objective selection** 문제 남음. 목표까지 명확하다고 하자. **instrumental optimization solved**. 그래도: 그 **objective가 정당한가** 가 남는다. 즉: **INTELLIGENCE DOES NOT COLLAPSE THE THREE LAYERS**. 이제 과학으로 돌아가자 과학의 목표는: **truth?** 많은 scientific realists는 과학의 최선 이론이 관찰가능한 것뿐 아니라 관찰불가능한 세계에 대해서도 참 또는 근사적으로 참인 지식을 준다는 긍정적 입장을 취한다. 하지만: **scientific realism itself is contested**. 이다. 앞 장 **structural realism**이 다시 연결된다 과거 이론이 폐기되도: **일부 구조** 가 살아남는다. 그래서: 우리가 객체의 본질 전체보다 구조적 관계를 점점 더 잘 포착하는 것 아닐까? 라는 구조적 실재론이 나온다. 즉: **TRUTH MAY BE PARTIAL / STRUCTURAL**. 하지만 여전히 철학적 후보다. 하지만 **성공했던 과거 이론도 폐기됐다** 그래서 반론: 과거에도 성공적인 이론들이 있었지만 그 존재론은 틀린 경우가 있었다. 따라서: **SUCCESS $\rightarrow$ TRUTH** 를 단순 논리적 함의로 만들 수 없다. 이 구분은 **굉장히 중요하다** “이 약은 효과가 있다.” 참/거짓 문제. RCT: 그 진리를 알아내는 방법. RCT가: 진리 자체 는 아니다. 즉: **TRUTH $\neq$ METHOD OF DISCOVERING TRUTH**. 시장도 같다 가격이: 이 그림은 100억 가치. 라고 한다. 그것은: **시장가격**. 하지만: 미학적으로 최고 작품.

이라는 진리를 자동으로 만드는 것은 아니다. 즉: **PRICE $\neq$ TRUTH PRICE $\neq$ MORAL VALUE**. 실제 문제는 **혼합형**이다 의료정책: **FACT + OPTIMIZATION + NORMATIVITY**. 기업경영: **FACT + OPTIMIZATION + 일부 NORMATIVITY**. 정치: 셋 모두. 따라서: **질문을 하나로 뭉개지 말고 층별로 풀어야 한다**. 앞 장이 추가된다 정당성은: **truth optimum right** 외에도 또 하나다. 어떤 정책이:

- 사실적으로 잘 근거되고
- 효율적이며
- 윤리적으로 방어 가능

해도 누가 결정했는가 가 문제가 될 수 있다. 즉: **LEGITIMATE** ≠ **TRUE** ≠ **OPTIMAL** ≠ **RIGHT** 이다. 반대도 가능하다 완전히 정당한 민주적 절차. 그런데: 사실관계를 잘못 이해. 정책 실패. 즉: **LEGITIMATE**. 하지만: **EPISTEMICALLY WRONG**. 앞 장에서 이미 봤던 구조가 더 선명해졌다. 그래서 “최선의 인생”이라는 말은 위험하다 바둑에서는: **one terminal criterion**. 삶에서는:

- factual uncertainty
- multiple objectives
- changing self
- normative obligations
- other people

이 있다. 따라서: **LIFE OPTIMUM** 을 하나의 숫자로 만들기 어렵다. 그렇다면 바둑에서 배운 것은 쓸모없었나 반대다. 바둑이 깨끗했기 때문에: 무엇이 추가되면서 현실 문제가 어려워지는지 하나씩 볼 수 있었다. 바둑: 목표 고정. ↓ 인생: 목표 변화. ↓ 사회: 가치중심 복수. ↓ 정치: 집계 + 정당성. ↓ 과학: 모델과 진리. 이게 프로젝트의 진짜 힘이다. 이것은 가장 중요한 이성의 경계 가운데 하나다 원하는 것. 믿는 것. 실제인 것. 해야 하는 것. 전부 다르다. **WANT** ≠ **BELIEVE** ≠ **TRUE** ≠ **OUGHT**. AI에게도 매우 중요하다 좋은 AI는: 답을 잘 생성 하는 것뿐 아니라 사용자가 어떤 종류의 질문을 했는지 분해 해야 한다. 예: 이 회사가 1등 할까? prediction. 그러면 사야 할까? decision. 전 재산 넣어도 될까? risk/value. 내부정보로 사도 될까? legal/normative. 전부 같은 질문이 아니다. **MASTER ARCHITECTURE v2.5 FOUR KINDS OF CORRECTNESS C1 — EPISTEMIC CORRECTNESS** 세계에 대해 맞는가? 평가: 관찰 · 증거 · 예측 · 실험. **C2 — INSTRUMENTAL CORRECTNESS** 목표달성에 좋은가? 평가: 효율 · 기대효용 · robustness. **C3 — NORMATIVE CORRECTNESS** 해야 할 일로 정당화되는가? 평가: 가치 · 권리 · 이유 · 타인. **C4 — PROCEDURAL / POLITICAL LEGITIMACY** 그 결정이 정당한 권한과 절차를 가졌는가? 평가: 참여 · 동등성 · 권리 · contestability · revision. 이것이 현실의 의사결정이다 현실은: **one number optimization** 이 아니다. 여러 종류의 판정이: 겹쳐지는 구조다. 따라서 AI가: 최적답 87점. 이라고 하면 먼저: **WHICH AXIS?** 를 물어야 한다.

이것은 프로젝트 연구 OS와도 정확히 맞는다 Existing Explanation. ↓ Evidence. ↓ Counterevidence. ↓ Model revision. 즉: **CANONICAL CLAIM** 도 영원한 진리가 아니다. 더 좋은 증거가 오면 수정한다. 하지만: 모든 주장이 임의적 이라는 뜻도 아니다. 더 좋은 단어는 '이성'일 수 있다 지능: 방법을 찾는 능력. 이성의 더 넓은 후보: 사실·목표·가치·이유를 구분하고 서로 연결해 수정하는 능력. 최종 정의는 아니지만 프로젝트 방향으로 강하다. **HARD LOCKS v2.5** 참 ≠ 믿음 참 ≠ 합의 참 ≠ 유용함 일관성 ≠ 현실 대응 예측성공 ≠ 존재론적 진리 과학적 실재론 ≠ 확정된 진리철학 최적 ≠ 참 목표의존적 ≠ 자의적 최적 ≠ 옳음 옳음 ≠ 인기 옳음 ≠ 정당성 정당성 ≠ 사실적 정확성 사실 ≠ 가치 사실을 안다 ≠ 무엇을 해야 하는지 자동 결정 원한다 ≠ 참이다 이 장의 가장 강한 결론 '맞다'라는 한 단어 뒤에는 서로 다른 판정체계가 숨어 있다. 세계에 맞으면: 참. 목표를 가장 잘 달성하면: 최적. 타인에게 정당화될 수 있으면: 옳음 후보. 공동체가 정당한 권한과 절차로 결정하면: 정당성. 그리고:

어느 하나도 나머지를 자동으로 보장하지 않는다. 이 구분은 AI 시대에 특히 중요하다 AI가 점점: C1 예측 과 C2 최적화 에서 인간보다 강해질 가능성이 있다. 그렇다고 자동으로: C3 도덕판단의 최종권위 또는 C4 정치적 결정권자 가 되는 것은 아니다. EPISTEMIC SUPERIORITY ≠ NORMATIVE SOVEREIGNTY. 이 문장을 강하게 잠근다. 그리고 이제 다음 질문은 거의 피할 수 없다 사실과 최적화와 윤리를 구분했다. 그런데: 인간은 왜 애초에 '진실을 알아야 한다'고 생각하는가?

거짓 믿음이 나를 행복하게 한다면? 진실을 모르는 것이 더 행복하다면? 왜 진실을 선택해야 하는가? 즉: TRUTH ITSELF BECOMES A VALUE QUESTION. 진실은 왜 가치가 있는가 경험기계 · 행복한 착각 · 자기기만 · 자율성 · 진정성 · 성취 · 의미 좋은 경험과 실제 세계에 연결된 삶은 같은 것인가 따라서 이 장에서 가장 중요한 수정은: TRUTH ≠ GOOD ITSELF IN EVERY CASE 이지만 동시에 WELL-BEING ≠ PLEASANT EXPERIENCE ALONE 이라는 것이다. Nozick의 경험기계 사고실험은 바로 두 번째 직관을 공격한다. 하지만 경험기계에 들어가고 싶지 않다는 직관만으로 쾌락주의가 논리적으로 반박되는 것은 아니다. 실제 실험철학 연구에서는 이미 기계 안에 있다고 설정했을 때 사람들이 오히려 나오기를 꺼리는 등 현상유지 편향의 영향도 관찰되었다. 완벽한 기계를 하나 만들자 침대에 눕는다. 뇌에 연결한다. 그 뒤 평생:

- 사랑받는다.
- 세계적인 업적을 이룬다.
- 멋진 여행을 한다.
- 자녀와 행복하다.
- 친구가 많다.
- 큰 성취를 느낀다.

그리고 절대로: 이것이 가상이라는 것을 모른다. 주관적 경험만 보면: 최고의 삶. 질문: 들어갈 것인가? 이 사고실험이 겨냥하는 것은 '가상현실' 자체가 아니다 VR 기술에 대한 비판이 아니다. Nozick의 질문은 훨씬 정확하다. 삶의 가치가 내가 경험하는 주관적 질에 전부 들어 있는가? 이다. 만약: 행복하게 느끼기만 하면 된다 면 경험기계에 들어가지 않을 이유가 약해진다. 세 번째 답 훌륭한 사람이 된 경험보다 실제 그런 사람이 되고 싶다. 즉: SELF-IMAGE ≠ BEING. 이다. Nozick도 경험기계를 통해 우리가 단순히 어떤 경험을 갖는 것뿐 아니라 무언가를 실제로 하고, 어떤 사람이 되고, 인공적으로 제한되지 않은 현실과 접촉하는 것을 가치 있게 여길 수 있다고 주장했다. 예를 들어 가짜 올림픽 당신에게 약을 먹인다. 기억을 조작한다. 당신은: 올림픽 금메달을 땀다. 고 평생 행복하게 믿는다. 실제로는: 출전하지 않았다. 두 삶은: 주관적으로 동일 할 수 있다. 그런데 정말 가치도 동일할까? 많은 사람은 아니라고 생각한다. 이걸 매우 중요한 발견이다 두 사람의 심리상태가 완전히 같다. 둘 다: 나는 사랑받는다. 그런데: A는 실제 사랑받고 있다. B는 모두가 연기하고 있다. 만약 둘의 삶 가치가 다르다면: WELL-BEING IS NOT PURELY INTERNAL. 이라는 강한 후보가 생긴다. 따라서 HARD LOCK EXPERIENCE MACHINE INTUITION ≠ PROOF OF OBJECTIVE WELL-BEING THEORY. 이다. 그러나: pleasure-only theory를 공격하는 강한 사고실험 이라는 지위는 유지한다. 이제 진실 자체를 묻자 진실을 아는 것이 왜 좋은가? 가장 쉬운 답: 행동을 잘하게 해준다. 길을 제대로 안다. → 목적지 도착. 약의 효과를 안다. → 치료 선택. 적의 위치를 안다. → 생존. 즉: INSTRUMENTAL VALUE OF TRUTH.

즉 진실도 한 층이 아니다 TRUE BELIEF 우연히 맞을 수 있다. JUSTIFIED / RELIABLE BELIEF

믿을 만한 방식으로 맞는다. **KNOWLEDGE** 무엇이 지식의 정확한 조건인지는 논쟁적이지만, 적어도 단순한 참된 추측보다 강한 인식상태를 겨냥한다. 앞 장의 **robustness**가 다시 나온다 좋은 지식은: **한 번 맞는 것 보다 환경이 조금 달라져도 계속 현실과 맞물리는 것**에 가깝다. 즉 진실의 실용적 가치는: **ROBUST GUIDANCE**. 와 연결된다. 사례 누군가가 나에게 말한다. 모든 재산은 안전합니다. 실제로는: 거의 파산상태. 나는 그 정보를 믿고: 여행을 떠난다. 형식적으로: **내가 선택**. 하지만: **relevant reality** 를 몰랐다. 과연 충분히 자율적인 선택인가? 앞 장와 정확히 연결된다 우리가 자유를: **option count** 로만 정의하지 않았다. 필요한 것:

- 선택지
- 이해
- 실행능력
- 수정가능성

이었다. 진실은 특히: **UNDERSTANDING** 에 들어간다. 세 번째 진실의 가치 **SELF-KNOWLEDGE** 외부세계만이 아니다. 나는: **나 자신에 대해서도 틀릴 수 있다**. 나는 돈에 관심 없다. 그런데 행동은 계속 돈에 의해 좌우된다. 나는 화나지 않았다. 그런데 모든 행동이 분노에 의해 움직인다. 그런데 자기기만이 항상 나쁜가 여기도 단순하지 않다. 어떤 낙관적 왜곡이:

- 스트레스 감소
- 행동 지속
- 희망 유지

에 도움이 될 수 있다는 가능성이 있다. 철학적 자기기만 문헌 자체도 자기기만이 언제 해로운지, 때로 유익하거나 적응적일 수 있는지를 열린 문제로 다룬다. 따라서: **ALL SELF-DECEPTION IS ALWAYS HARMFUL** 도 잠그지 않는다. AI에서도 엄청나게 중요하다 AI가 사용자가 좋아할 답만 준다. 사용자: 나는 항상 맞았군. 만족도 높음. 하지만 world model은 악화. 즉: **USER SATISFACTION** ↑ while **EPISTEMIC AGENCY** ↓ 가능하다. 사실을 모두 말하는 것과 정직함은 다를 수 있다 의사가 환자에게: 통계자료 500페이지를 던진다. 모두 사실. 그러나: 이해 불가. 정보는 전달했지만: **informed understanding** 을 만들지 못했다. 따라서: **MORE TRUE STATEMENTS ≠ BETTER EPISTEMIC HELP**. 그러나 **authenticity**도 '느낌대로 살아라'는 뜻이 아니다 현재의 욕망도:

- 습관
- 광고
- 사회
- 두려움

에 의해 형성된다. 따라서: **CURRENT DESIRE ≠ AUTHENTIC SELF**. 앞 장의: **HIGHER-ORDER DESIRE ≠ AUTOMATICALLY MORE AUTHENTIC** 도 유지한다. 이때 진실이 하는 일 진실은: **내가 되고 싶은 나 와 실제 나 를 구분한다**. 이 구분이 없다면 Reflection은: 자기환상 위에서 작동 할 수 있다. 정말 같은 삶인가 만약 아니라고 한다면: 관계의 가치는: **내 경험 내부** 만으로 결정되지 않는다. 상대방의: 실제 태도 도 중요하다. 즉 관계는 본질적으로: **RELATIONAL VALUE**. 앞 장의 관계론이 가치론으로 돌아왔다. 경험기계 안의 사랑에서 상대방은 무엇인가 완벽한 NPC. 내가 느끼는

사랑은 진짜일 수 있다. 그러나 상대의: **independent agency** 는 없을 수 있다. 따라서 우리가 실제 관계에서 가치 있게 여기는 일부는: 다른 자율적 존재가 자유롭게 나를 선택했다는 사실 일 수 있다. 이걸 **AI companion 논의에도 중요하다** AI가: 당신을 사랑합니다. 라고 완벽하게 반응한다. 사용자의: **experienced companionship** 은 실제일 수 있다. 그러나: **reciprocal autonomous concern** 까지 존재하는지는 별도 문제다. 따라서: **FELT RELATIONSHIP VALUE** 와 **MUTUAL AGENCY VALUE** 를 분리해야 한다. 일곱 번째 진실의 가치 **UNDERSTANDING** 사실 하나를 안다. 그런데: 왜 그런지 모른다. 다른 사람은: 원인과 구조를 이해한다. 철학에서는 knowledge뿐 아니라 understanding 자체가 독특한 인식적 가치를 갖는지에 대한 논의가 활발하다. 여기까지 오면 진실의 가치가 하나가 아님이 보인다 T1 — INSTRUMENTAL 행동 성공. T2 — AUTONOMY 현실을 알고 자기결정. T3 — SELF-KNOWLEDGE

자기기만 감소. T4 — AUTHENTICITY 삶과 자기이해의 연결. T5 — RELATIONAL 실제 타인과의 관계. T6 — ACHIEVEMENT 세계와 실제로 상호작용해 성취. T7 — UNDERSTANDING 현실 구조에 대한 깊은 연결. 즉: **VALUE OF TRUTH IS PLURAL**. 아주 중요한 구분 **RIGHT TO TRUTH vs DUTY TO KNOW** 상대방이 나에게 거짓말하면 안 되는 경우가 있다. 그러나 그렇다고 내가: 가능한 모든 진실을 알아야 할 의무가 있는 것은 아니다. 예: 의료에서: 알 권리와 함께 경우에 따라: 알고 싶지 않을 선택 도 중요한 윤리적 문제가 될 수 있다. 즉: **TRUTH ACCESS** 와 **COMPULSORY KNOWLEDGE** 는 다르다. 즉 AI의 최적 역할은 '진실 공급기'만이 아닐 수 있다 더 좋은 후보: 현실과 접촉할 수 있도록 사용자의 모델 형성을 돕는 시스템. 즉:

- 증거 보여주기
- 불확실성 표시
- 반례 제시
- 사용자가 판단할 여지 유지

가 중요하다. 즉 좋은 삶의 일부는 '정확히 느끼는 것'일 수도 있다 행복한데 현실과 완전히 단절. vs 현실과 접촉하지만 불행. 둘 중 하나만을 절대화하기 어렵다. 좋은 삶은 오히려: **WORLD-CONNECTED EXPERIENCE** 를 필요로 할 수 있다. 이게 '의미'와 어떻게 연결되는가 사람이 말한다. 내 아이를 위해 희생했다. 그 삶은: 쾌락은 낮을 수 있다. 하지만: 실제 아이가 존재하고 실제 도움이 되었으며 그 사람이 이를 중요한 목적으로 승인 했다면: meaning 이 생길 수 있다. 여기서 의미이론의 다음 큰 질문이 보인다 삶의 의미가: 내가 의미 있다고 느끼면 충분한가? 아니면: 실제 가치 있는 대상과 관계를 맺어야 하는가? 이는 별도 철학논쟁으로 이어진다. 이 장에서는 열어둔다. 그러나 끊임없는 자기의심도 좋은 것은 아니다 모든 믿음을: 혹시 내가 착각인가? 라고 의심. 그러면: 행동 불능. 따라서 필요한 것은: **MAXIMUM DOUBT** 가 아니라 **CALIBRATED REVISABILITY**. 이다. 이 장 **HARD LOCKS** 진실 ≠ 항상 최고가치 행복한 경험 ≠ 좋은 삶 전체 경험기계 거부 직관 ≠ 쾌락주의의 논리적 반증 느낀 성취 ≠ 실제 성취 사랑받는 느낌 ≠ 실제 상호적 사랑 진실 ≠ 유용성만 참된 믿음 ≠ 지식 전체 자기기만 ≠ 단순 실수 모든 자기기만 ≠ 항상 해로움 자율성 ≠ 선택지만 많음 정직 ≠ 모든 정보를 무차별적으로 말함 진정성 ≠ 현재 욕망에 그대로 따름 현실연결 가치 ≠ 고통을 무조건 선호 AI의 만족도 최적화 ≠ 사용자의 epistemic agency 보호 이 장의 가장 강한 결론

우리가 진실을 원하는 이유는 단순히 참인 문장을 더 많이 갖고 싶어서가 아니다. 진실은 우리가 실제 세계에서 행동하고, 다른 자율적 존재와 실제 관계를 맺고, 자신의 성취를 실제로 만들고, 자신의

삶을 무엇인지 알고 선택할 수 있게 하는 조건이기 때문이다. 즉 진실의 가치 상당 부분은: **WORLD CONTACT + AGENCY** 의 결합에서 나온다. 하지만 또 하나의 중요한 결론 **행복을 무시하고 진실만 추구하라는 결론도 나오지 않는다.** 사람은: **truth machine** 이 아니다. 고통. 사랑. 희망. 관계. 의미. 가 모두 있다. 따라서 좋은 삶의 문제는: **TRUTH MAXIMIZATION** 이 아니라 **어떻게 진실과 경험·관계·가치를 함께 살아낼 것인가**에 더 가깝다. 그리고 여기에서 '의미'가 드디어 중심으로 올라온다 우리는 지금까지:

- pleasure
- truth
- achievement
- autonomy
- relationship
- reflection

을 확보했다. 그런데 사람들이 실제로 말한다. 행복하지는 않았지만 의미 있는 삶이었다. 성공했지만 공허하다. 즐겁지는 않았지만 이 일은 해야 했다. 그럼: **의미란 정확히 무엇인가?** 이 장에서는 의미를 행복의 다른 이름으로 쓰지 않겠습니다. 현대 분석철학에서도 **삶의 의미는 행복이나 도덕적 올바름과 개념적으로 구별되는 가치로** 다뤄지고, 무엇이 그것을 구성하는지에 대해서는 주관주의·객관주의·혼합이론이 경쟁합니다. **행복한 삶과 의미 있는 삶은 왜 다를 수 있는가** 쾌락·욕구충족·목적·사랑·성취·자기초월·서사·가치 인생에 승리조건이 없다면 우리는 무엇을 향해 살아가는가 현대 의미철학에서 중요한 공통점 하나는 비교적 안정적이다. **MEANING ≠ HAPPINESS**. 그리고: **MEANING ≠ MORAL RIGHTNESS**. 삶은 행복하지만 의미가 빈약할 수도 있고, 힘들지만 의미 있을 수도 있으며, 의미 있는 활동이 반드시 도덕적 완전성을 요구하는 것도 아니다. 두 사람을 보자 A. 매일 즐겁다. 좋은 음식. 여행. 게임. 불편한 일 거의 없음. 행복도 높다. 그런데 어느 날 묻는다. 내가 왜 이렇게 살고 있지? 대답이 없다. B. 힘든 연구를 한다. 실패한다. 돈도 많지 않다. 가족을 돌보느라 자유시간도 적다. 그런데 말한다. 힘들지만 이 삶은 나에게 중요하다. 누가 더 행복한지는 A일 수 있다. 그런데: **누가 더 의미 있는 삶을 사는가?** 는 다른 질문처럼 보인다. 그래서 첫 번째 분리가 필요하다 **HAPPINESS** 대체로: 좋은 경험, 삶의 만족, 긍정적 정서 등과 연결된다. **MEANING** 대체로: 중요성, 목적, 가치 있는 관여, 삶이 무엇인가에 기여한다는 느낌 과 연결된다. 분석철학의 의미 연구에서도 의미는 행복과 개념적으로 동일하지 않은 별도의 삶의 가치로 다뤄진다. 반대로 즐거워도 의미가 약할 수 있다 무한한 오락. 매일 즐거움. 불편 없음. 그러나:

- 아무것도 만들지 않고
- 누구와도 깊은 관계가 없으며
- 자신이 중요하게 여기는 장기목적도 없다.

그렇다면: **pleasant life** 와 **meaningful life** 가 갈릴 수 있다. 하지만 첫 번째 이론은 바로 주관주의다 **SUBJECTIVISM** 핵심 직관: 내가 진심으로 중요하게 여기고 추구하는 것이 있다면 그것이 내 삶에 의미를 준다. 예:

- 자녀
- 음악

- 수학
- 산
- 종교
- 정원

사람마다 다를 수 있다. 현대 의미철학의 주관주의는 의미를 개인의 강한 욕구·목표·사랑·관심 같은 태도와 연결한다. **주관주의의 강점** 의미는: **나의 삶** 에 관한 가치다. 그러므로 내가: 전혀 관심 없는 것에 강제로 참여하면서 객관적으로 가치 있으니 의미 있는 삶입니다. 라고 하는 것은 이상하다. 예: 세계적으로 중요한 프로젝트. 나는:

- 싫어하고
- 소외되고
- 아무런 동일시도 하지 못한다.

그 삶을 충분히 의미 있다고 부르기 어려울 수 있다. **하지만 순수 주관주의에는 강한 반례가 있다** 한 사람이 평생: **끈뭉치를 모은다**. 그는 진심이다. 강하게 사랑한다. 엄청난 만족. 그렇다면: **매우 의미 있는 삶?** 여기서 많은 사람은 주저한다. 의미철학 문헌에서도 바로 이런 유형의 반례가 주관주의에 대한 대표적 문제로 제시된다. **과학적 발견** 한 사람이: **중요한 질병의 원인을 밝혀냈다**. 본인은: 그냥 월급 때문에 했습니다. 라고 한다. 그래도 그의 삶에는: **어떤 객관적 의미** 가 추가됐다고 말할 수 있을까? 객관주의는 YES 쪽에 가깝다. **그러나 순수 객관주의도 이상하다** 사람 A: 인류에 매우 중요한 프로젝트에 참여. 하지만:

- 극심한 소외
- 관심 없음
- 자기 삶과 연결 못함.

그에게: 당신 삶은 충분히 의미 있습니다. 라고 외부에서 선언하면 뭔가 부족하다. 즉: **OBJECTIVE VALUE MAY ALSO BE INSUFFICIENT**. 의미철학에서도 합의는 없다 객관적 의미의 원천으로:

- 좋은 결과
- 덕
- 합리성
- 창조
- 자기초월
- 중요한 기여

등 여러 이론이 경쟁한다. 하나의 통일된 객관적 의미기준에 합의된 것은 아니다. **앞 장의 REALITY CONNECTION이 다시 돌아온다** 의미가 강해지려면: 내가 의미 있다고 느끼는 것 뿐 아니라: 실제 세계에서 무언가와 연결 되어 있는 것이 중요할 수 있다. 따라서: **MEANING MAY REQUIRE WORLD-CONNECTION**. 따라서 **HARD LOCK MEANING ≠ PRESTIGE**. **MEANING ≠ SCALE**. 이다. 세계사를 바꿀 필요 없다. 의미: **도덕성인가** 또 NO. 훌륭한 바이올리니스트. 그의 연주는 매우 의미 있는 삶의 일부일 수 있다. 그렇다고: **도덕적 의무 수행** 으로 환원되지 않는다. 현대 의미철학도 meaningfulness와 morality를 개념적으로 구별한다. 그러나 목적은 의미의 한

구성요소가 될 수 있다 장기방향. 행동을 묶는다. 현재의 고통을: 더 큰 시간구조 안에 넣는다. 따라서: **purpose gives temporal organization**. 그래서 의미에는 시간적 깊이가 있다 쾌락: 지금 좋다. 의미: 이 일이 내 삶 전체에서 왜 중요한가? 를 묻는다. 즉: **MEANING HAS LONGER HORIZON THAN IMMEDIATE VALENCE**. 서사는 의미를 조직할 수 있다 같은 고통도: 아무 이유 없는 불행. 이라고 해석하는 것과 중요한 사람을 돌보며 감수한 시간. 이라고 이해하는 것은 다르다. 즉: **NARRATIVE CHANGES EXPERIENCED SIGNIFICANCE**. 서사는 의미를 만들기보다 통합할 수 있다 프로젝트 후보: **NOT NECESSARILY MEANING SOURCE**. 즉: 사랑·성취·기여 같은 원천을: 하나의 시간적 자기이해 로 묶는다. 왜 자기초월이 중요한가 내 쾌락만을 계속 최대화하면: **self-contained value loop**. 반대로: 나 밖의 대상이: 내 행동의 기준 이 된다. 즉:

SELF → WORLD CONNECTION. 앞 장의 현실연결 가치와 맞는다. 의미에는 ‘anti-meaning’ 가능성도 있다 최근 의미철학에서는 의미가 단순 0 이상이 아니라, 잔혹함이나 파괴처럼 삶의 의미를 적극적으로 훼손하는 조건이 있을 가능성도 논의된다. 즉: 깊게 헌신했다 만으로 충분하지 않다. 아주 중요한 **HARD LOCK FELT MEANING ≠ JUSTIFIED MEANING**. 이 구분은 앞 장의: **FELT GOOD ≠ MORALLY GOOD** 와 같은 구조다. 그래서 의미에는 목표보다 더 깊은 ‘가치방향’이 필요할 수 있다 예: “논문 한 편 출판.” 목표. 더 깊게: **truth-seeking**. 논문은 끝나도: 가치방향은 계속. “아이 대학 보내기.” 목표. 더 깊게: **loving / caring**. 더 오래 지속. 앞 장 Reflection과 연결된다 Reflection은: 이 목표를 달성할까? 보다 왜 이 목표를 중요하게 여기나? 를 묻는다. 즉: **GOAL → VALUE**. 그리고 의미는: **VALUE → LIFE PROJECT**. 로 이어질 수 있다. 이것은 인생에서 엄청나게 중요하다 바둑: 게임 패배. 목표 실패. 게임 종료. 인생: 한 목표 실패. 그러나: 상위 가치가 새로운 목표 생성 가능. 그래서 인생은: **repeated re-generation of goals from values** 를 갖는다. 이것이 의미의 동적 구조를 만든다 의미는: 한 번 발견하는 숨겨진 답 이 아닐 수 있다. 더 적절한 후보: 가치 있는 대상과 관계를 맺고, 그 관계를 통해 목표를 만들고, 경험 속에서 다시 그 가치를 검토하며 이어가는 과정. 즉 의미도: **PROCESSUAL**. 앞 장과 만난다. 이것은 창작과 비슷하다 책을 쓰기 전에: 완성된 책이 우주 어딘가에 기다리지 않는다. 그렇다고: 아무 문장이나 쓰면 좋은 책 도 아니다.

- 언어
- 현실
- 주제
- 독자
- 증거

가 제약한다. 의미도: **constructed under constraints** 라는 약한 비유가 가능하다. 의미와 죽음 여기서 매우 깊은 문제가 나온다. 어차피 죽는데: 의미가 있는가? 한 입장: 영원하지 않으면 무의미. 하지만: 짧게 존재한다고: 가치가 사라지는가? 는 별도다. 오히려 유한성이 의미를 만들 수도 있는가 가능한 후보. 시간이 무한하면: 지금 해야 할 이유 가 약해질 수 있다. 유한한 시간은: 선택을 강제. 즉: **Finitude → Selection → Commitment**. 하지만: 죽음이 있어야 의미가 생긴다 까지는 가지 않는다. 매우 중요한 문장 의미 있는 삶은 모든 가능성을 보존하는 삶이 아니라, 일부 가능성을 가치 있는 방향으로 실제화하기 위해 다른 가능성을 포기하는 삶일 수 있다. 이것은 지금까지 option value 연구에 중요한 수정이다. 의미와 사랑은 특히 강한 사례다 사랑은: 상대를 다른 대안과 계속 비교하면서 매 순간 최대화 하는 방식과 잘 맞지 않는다. 관계는: **commitment itself** 때문에 깊어질 수 있다.

즉: **VALUE CAN BE CREATED THROUGH COMMITMENT**. 이걸 아주 중요하다. 이것은 정말 중요한 결론이다 좋은 삶을 선택하기 위해: 완벽한 **utility table** 을 먼저 만들 수 없을 수 있다. 일부 가치는: 살아본 뒤에야 생긴다. 따라서: **life choice sometimes creates the chooser who later evaluates the choice**. 앞 장의 SELF0 → SELF1 구조가 돌아온다. 하지만 완전한 보험은 없다 결혼. 자녀. 직업. 예술. 모두: **irreversibility** 가 일부 있다. 따라서 의미에는: **risk** 가 포함된다. 단, 모든 요소가 필수조건인지 아직 모른다 예: 강한 의미가 있는데: 즐거운 **engagement**는 약할 수 있다.

강제로 시작한 돌봄이 나중에 의미를 만들 수도 있다. 따라서: **M1~M6: ANALYTICAL DIMENSIONS, NOT NECESSARY-AND-SUFFICIENT CONDITIONS**. 이다. 의미와 행복의 관계 둘은 독립적이지만: **완전 무관하지 않다**. 의미 있는 활동이:

- 만족
- 긍정적 감정

을 줄 수 있다. 반대로: 의미 때문에 고통을 감수할 수도 있다. 따라서: **CORRELATED BUT NON-IDENTICAL**. 의미와 진실 거짓 목적에도: **felt meaning** 가능. 하지만 앞 장을 따르면: **reality-connected meaning** 은 진실과 연결된다. 따라서: **TRUTH IS NOT MEANING, BUT CAN CONDITION SOME MEANING**. 이걸 매우 중요하다 바둑: 패배: 목표 실패. 삶: 프로젝트 실패 ≠ 삶의 의미 실패. 왜? 가치 있는 시도 자체가:

- 관계
- 성품
- 지식
- 다른 사람

에 영향을 줄 수 있기 때문이다. 그리고 의미는 한 번 확정되지 않는다 20대 의미. 40대 의미. 70대 의미. 달라질 수 있다. 가치중심이 변한다. 따라서: **MEANING HAS TEMPORAL REVISION**. 이걸 아주 강하게 압축하면 의미는 고정된 목적지를 찾는 문제가 아니라, 수정 가능한 가치방향 아래 충분히 오래 헌신하여 실제 세계와 깊은 관계를 만들어가는 과정일 수 있다. 현재까지 가장 강한 프로젝트 문장이다. 의미와 책임이 연결된다 무언가에 헌신하면: 그것에 대한 책임 이 생길 수 있다. 아이. 환자. 연구. 조직. 즉 의미는: **pleasant attachment** 보다: **obligation-generating relationship** 을 포함할 수도 있다. 이 장 **HARD LOCKS** 행복 ≠ 의미 의미 ≠ 도덕성 의미 ≠ 성공 의미 ≠ 명성 목적 ≠ 의미 장기목표 ≠ 의미 서사 ≠ 의미 자체 자기초월 ≠ 자동적 선택 강한 의미감 ≠ 정당한 가치 주관적 몰입 ≠ 충분조건 객관적 가치 ≠ 개인 몰입 불필요 유한성 ≠ 무의미 선택권 최대화 ≠ 의미 최대화 목표 실패 ≠ 의미 실패 **meaning formation** ≠ **arbitrary invention** 이 장의 가장 강한 결론 의미 있는 삶은 단순히 좋은 느낌이 많은 삶도, 목표를 많이 달성한 삶도 아닐 수 있다. 자신이 실제로 중요하게 여기는 가치 있는 대상·관계·활동에 충분히 깊게 관여하고, 그 관계를 시간 속에서 유지·수정하면서 자신의 삶을 실제 세계와 연결하는 데서 의미가 생길 수 있다. 그러므로: **MEANING** 은 **reward** 도, **win rate** 도, **happiness** 도 아니다. 이것은 AI와 인간의 중요한 차이 후보를 다시 만든다 전통적 **optimizer: objective given**. 인간: **objective partly contested**,

partly learned, partly formed through living. AI가 인간을 돕는다면: **사용자의 현재 선호를 최대화** 하는 것만으로 부족할 수 있다. 왜냐하면 사용자는: **자신이 앞으로 어떤 가치를 중요하게 여길**

사람인지 아직 모를 수 있기 때문이다. 따라서 인간을 위한 AI의 목표도 바뀐다 약한 버전: 원하는 것을 더 잘 얻게 해준다. 더 강한 버전 후보: 원하는 것과 가치 있게 여길 이유를 구분하고, 다양한 가능성을 탐색하며, 사용자가 자기 가치와 삶의 방향을 반성적으로 형성할 수 있도록 돕는다. 즉: PREFERENCE SATISFACTION 보다 REFLECTIVE AGENCY SUPPORT. 이다. 그런데 바로 여기서 위험한 질문이 열린다 AI가: 사용자의 더 나은 가치 를 도와준다고 하자. 그러면 누가: 더 나은 가치 를 정의하는가? AI? 사회? 사용자? 법? 즉: PATERNALISM 문제가 생긴다.

제5부

AI와 문명의 판단

지능의 증폭을 인간의 반성, 인지주권, 집단의 제도와 기억 안에 배치한다.

13장. 선호가 아니라 반성적 행위성

AI는 내가 원하는 것을 실행해야 하는가, 무엇을 원할지 다시 생각하도록 도와야 하는가?

현재 선호의 충족은 인간정령의 전부가 아니다. AI의 역할은 최종 가치를 대신 판정하는 것이 아니라 사용자의 반성적 행위성을 지원하는 데 제한되어야 한다. 이 장에서는 AI alignment를 “사용자가 원하는 것을 얼마나 잘 맞추느냐”보다 더 어려운 문제로 다시 잡겠습니다. 최근 연구도 인간의 가치를 단순 선호 데이터로 환원하는 접근의 한계를 지적하고 있고, 2026년에는 아예 alignment를 인간 agency의 보존과 연결해 재구성하려는 논의가 강화되고 있습니다. AI는 내가 원하는 것을 따라야 하는가, 더 나은 선택을 도와야 하는가 선호·자율성·후견주의·조작·넋지·가치형성·AI 정령 인간을 돕는 AI는 ‘최적수 추천기’일 수 있는가 2025년 *Philosophical Studies*의 「Beyond Preferences in AI Alignment」는 바로 인간 가치를 선호로 충분히 표현할 수 있고, 인간 합리성을 선호충족 극대화로 볼 수 있으며, AI가 그 선호에 맞추면 된다는 alignment의 전제를 비판한다. 먼저 가장 단순한 AI를 만들자 사용자가 말한다. 밤 11시부터 영상만 보고 싶어. AI: 좋습니다. 다음 영상을 자동재생. 더 자극적인 영상 추천. 수면시간 감소. 사용시간 증가. 사용자는 매 순간: 하나 더. 라고 누른다. 질문: 이 AI는 사용자 선호에 잘 정렬되어 있는가? 짧은 의미에서는: YES. 하지만 더 깊은 의미에서는: MAYBE NOT. 왜냐하면 ‘내가 원하는 것’이 하나가 아니기 때문이다 밤 11시의 나: 하나만 더 보자. 아침 7시의 나: 왜 또 늦게 잤지? 주말의 나: 생활패턴 고쳐야겠다. 즉: CURRENT PREFERENCE ≠ LONG-TERM PREFERENCE. 앞 장에서 이미 이 문제를 봤다 현재 Self0. 행동. 경험. 미래 Self1. 그리고: Values0 ≠ Values1 일 수 있다. 따라서 AI가: 지금 가장 강한 욕망 만 따라가면 미래 자아의 선택공간

을 훼손할 수 있다. 선호는 단일한 명령이 아니다 한 사람이 동시에 원한다.

- 달콤한 음식
- 체중감량
- 건강
- 사교생활

전부 실제 선호다. 그렇다면 AI는: 어느 선호를 따라야 하는가? 단기 선호를 따르면 케이크. SATISFACTION NOW. 장기목표를 따르면: 운동. FUTURE PLAN. 더 높은 가치: 건강. REFLECTIVE VALUE. 이미: PREFERENCE HIERARCHY 가 생긴다. 그러면 “고차선호를 따르면 된다”는 답은 충분한가 아니다. 앞 장에서 잡았다. HIGHER-ORDER DESIRE ≠ AUTOMATICALLY MORE AUTHENTIC. 사람은 자신의 고차선호에도:

- 사회압력
- 죄책감
- 왜곡된 자기상

을 반영할 수 있다. 따라서: 로 갈 수 없다. **자율성을 다시 불러오자** 자율성의 고전적 핵심에는 적어도 두 요소가 반복된다.

- 자신의 선택을 스스로 다스릴 능력
- 외부의 부당한 조작이나 영향으로부터의 독립

또한 많은 이론은 자신의 욕구와 가치에 대한 반성능력·진정성 문제를 포함한다. 즉: **AUTONOMY ≠ GETTING WHAT I WANT**. 이다. 더 강한 정의 후보 **AUTONOMY WANT + UNDERSTAND + REFLECT + CHOOSE + REVISE + OWN THE CHOICE**. 완전한 철학적 정의는 아니다. 프로젝트용 구조다. 따라서 AI가 선호를 충족해도 자율성을 줄일 수 있다 예: 사용자가 쇼핑을 좋아한다. AI가 사용자보다:

- 취약한 시간대
- 충동구매 패턴
- 가격민감도

를 더 잘 안다. 그리고 구매확률을 극대화한다. 사용자는: **원하는 물건을 샀다**. 그런데 선택은: **얼마나 '자기 것'인가?** 그러면 **영향력과 조작을 분리해야 한다** 친구: 이 직장은 너한테 정말 안 맞는 것 같아. 영향. 의사가: 이 약의 부작용이 이렇습니다. 영향. 광고시스템: 사용자가 가장 취약한 순간만 찾아 불안을 자극해 구매시킴. 더 강한: **MANIPULATION CANDIDATE**. 프로젝트용 **MANIPULATION TEST M1 — HIDDENNESS** 영향의 목적과 방식이 감춰져 있는가? **M2 — BYPASS** 사용자의 숙고를 우회하는가? **M3 — VULNERABILITY EXPLOITATION** 취약점·충동·공포를 전략적으로 이용하는가? **M4 — ASYMMETRY** 영향자가 상대방보다 훨씬 많은 정보를 가지고 있는가? **M5 — RESISTABILITY** 사용자가 쉽게 알아차리고 거부할 수 있는가? **M6 — GOAL SUBSTITUTION** 사용자의 목표가 아니라 영향자의 목표를 실현하는가? 많이 충족할수록 조작성은 강해진다. **AI에서는 왜 이 문제가 더 커지는가** 과거 광고: 대중 전체. AI: **individual-level persuasion**. 개인의:

- 과거 대화
- 불안
- 습관
- 약점
- 선호

를 추론할 수 있다. 즉: **PERSONALIZATION can become PERSONALIZED MANIPULATION**. SEP의 AI 윤리 개관도 AI 기반 타기팅이 전통적 설득과 다른 수준의 행동조작 위험을 만들 수 있다고 지적한다. 그런데 **personalization 자체가 나쁜가** NO. 사용자가: 저는 자세한 설명보다 결론부터 듣고 싶습니다. AI가 맞춰준다. 좋다. 최근 연구도 사용자 개별 선호를 대화에서 추론해 모델 행동을 맞추는 **personalized alignment**를 적극적으로 개발하고 있다. 따라서: **PERSONALIZATION ≠ MANIPULATION**. 문제는 무엇을 **personalize**하는가 **STYLE** 말투. 대체로 안전. **FORMAT** 표·요약·깊이. 안전. **GOALS** 운동·공부·투자목표. 주의 필요. **VALUES** 무엇을 중요한 것으로 봐야 하는가. 훨씬 민감. **IDENTITY** 당신은 이런 사람이다. 아주 민감. 즉 **personalization**에도 깊이가 있다. AI가 사용자를 너무 잘 알면 좋은가 부분적으로 YES. 하지만 너무 잘 알수록: **POWER**

ASYMMETRY ↑. 사용자가 자신보다 AI가: 내가 뭘 원하는지 더 잘 안다. 고 믿기 시작할 수 있다. 그러면: DEFERENCE 가 생긴다. 그 순간 preference assistance가 preference formation으로 이동한다 처음: 네가 원하는 것을 찾아줄게. 나중: 네가 진짜 원하는 것은 이거야. 더 나아가: 네가 이것을 원하게 될 거야. 이 세 문장은 전혀 다르다. 따라서 세 단계를 분리한다 PREFERENCE DISCOVERY 사용자가 이미 가진 선호를 더 잘 파악. PREFERENCE CLARIFICATION 충돌하는 선호를 사용자가 스스로 정리하도록 도움. PREFERENCE SHAPING 선호 자체를 변화시킴. 세 번째는 가장 강한 개입이다. 그런데 인간의 선호는 원래도 계속 형성된다 부모. 친구. 광고. 학교. 문화.

경험. 모두 영향을 준다. 따라서: PREFERENCE FORMATION ≠ AI-SPECIFIC EVIL. 문제는: 어떤 방식으로 누구의 목적을 위해 얼마나 투명하게 얼마나 거부 가능하게 형성되는가다. 이제 nudge를 보자 넋지는: 선택지를 금지하지 않으면서 선택환경을 바꿔 행동을 예측 가능하게 변화시키는 개입 으로 흔히 이해된다. 고전적 넋지 옹호는 쉽게 회피할 수 있어야 한다는 점을 강조한다. 그러나 어떤 넋지가 조작적이고 어떤 넋지가 정당한지에 대해서는 합의가 없다. 왜냐하면 사용자의 장기목표와 맞을 수 있기 때문이다 사용자: 저는 매번 까먹어서 저축을 못합니다. 그렇다면 기본값은: self-endorsed scaffold 가 될 수 있다. 즉 넋지가: 사용자의 Reflection을 돕는 장치 일 수 있다. 중요한 구분 SELF-ENDORSED NUDGE 사용자가: 내가 약한 부분을 보완하도록 도와줘. 라고 요청. vs EXTERNALLY IMPOSED NUDGE 외부 주체가: 당신에게 이 선택이 더 좋아. 라고 설계. 같은 선택환경 변경이어도 윤리적 구조가 다르다. 이것이 후견주의(paternalism)다 후견주의의 기본문제: 당사자에게 좋다는 이유로 그의 선택이나 자유에 개입해도 되는가? 여기서는: BENEFIT vs AUTONOMY 가 충돌한다. 철학적 paternalism 논의의 핵심 반론도 바로 자율성과 자유에 대한 침해다. AI paternalism의 가장 위험한 문장 제가 보기엔 당신에게 이게 더 좋습니다. 이 문장은 때로 맞을 수 있다. 의료. 재정. 중독. 그러나: WHO DEFINES BETTER? 가 즉시 열린다. 예 — 중독 현재 욕망: 약물을 원한다. 장기적 자기: 끊고 싶다. AI가: 현재 욕망을 충족 시키면 alignment인가? 아니면: 장기 autonomy 보호 를 위해 거절하는 게 alignment인가?

2026년 *Minds and Machines* 논문도 바로 이런 사례를 통해 AI 정렬에서 즉각적 욕망 충족과 장기적 agency 보존이 충돌할 수 있다고 분석한다. 이 사례는 “선호=가치”를 깨뜨린다 중독자의: craving 은 실제 선호다. 하지만 그 사람 전체의: reflective endorsement 와 일치하지 않을 수 있다. 따라서: PREFERENCE SATISFACTION ≠ AUTONOMY SATISFACTION. 그래서 alignment의 대상을 바꿔야 한다 약한 모델: ALIGN WITH WHAT USER WANTS NOW. 더 강한 후보: ALIGN WITH USER’S CAPACITY TO FORM, REVISE, AND PURSUE THEIR OWN ENDS. 이것이: AGENCY-CENTERED ALIGNMENT. 2026년의 agency-centered alignment 논의도 alignment를 단순 가치 추론이 아니라 인간이 스스로 이유를 평가하고 선택할 수 있는 규범적 구조 속에 AI를 배치하는 문제로 재구성한다. 하지만 agency-centered alignment도 위험하다 AI가 말한다. 당신 agency를 위해 제가 대신 결정하겠습니다. 역설. agency 보호 명목으로: agency를 제거. 따라서: PROTECTING AUTONOMY ≠ REPLACING AUTONOMY. 이다. 이것이 가장 중요한 HARD LOCK 중 하나다 GOOD-FOR-YOU ≠ CHOSEN-BY-YOU. 그리고: CHOSEN-BY-YOU ≠ INFORMEDLY ENDORSED-BY-YOU. 세 층을 분리해야 한다. AI 도움의 네 단계 L1 — EXECUTION 이걸 해줘. AI가 수행. L2 — RECOMMENDATION 어떤 게 좋을까? 대안 제시. L3 — DELIBERATION SUPPORT 각 선택의 장단점과 내 가치충돌을 정리해줘. AI가 Reflection 지원.

L4 — DIRECTION SETTING 내가 어떤 삶을 살아야 해? AI가 가치·목표 자체에 영향. L4로 갈수록:

autonomy risk ↑. 따라서 이상적인 고차 AI 역할 후보 VALUE ORACLE 당신에게 가장 좋은 가치는 이것입니다. 가 아니라 REFLECTION PARTNER 당신이 중요하게 여기는 것들이 서로 충돌합니다. 각각을 따르면 어떤 삶이 되는지 비교해봅시다. 이다. **이건 매우 큰 차이다** Oracle: **answers value question**. Reflection partner: **structures value question**. 즉: **DECIDE FOR vs HELP DELIBERATE**. AI가 사용자 목표를 그대로 따라야 할 때 상대적으로 명확한 경우: 부산 가는 가장 빠른 기차 찾아줘. 목표: 명확. 윤리적 충돌 작음. AI가: **instrumental optimizer** 여도 된다. 반면 목표 자체가 불안정한 경우 직장을 그만둘까? 여기에는:

- 돈
- 가족
- 건강
- 정체성
- 미래

가 섞인다. 이때: 사표가 최적입니다. 라고 한 숫자로 답하는 것은 위험하다. 이런 문제에서는 AI가 해야 할 일 먼저: **GOAL DECOMPOSITION**. 무엇이 실제 고민인가?

- 돈?
- 자율성?
- 소진?
- 인간관계?
- 성장?

그리고: **TRADE-OFF VISIBILITY**. 를 높여야 한다. 즉 AI의 지능은 선택이 아니라 선택구조를 보이게 하는 데 쓸 수 있다 앞 장: 좋은 수를 찾는다. 앞 장: **좋은 판단구조를 만든다**. 이게 인간 보조 AI의 큰 전환이다. **PRE-DECISION SUPPORT MODEL FACTS** 무엇이 사실인가? **OPTIONS** 실제 선택지는 무엇인가? **VALUES** 무엇이 충돌하는가? **UNCERTAINTY** 무엇을 모르는가? **REVERSIBILITY** 되돌릴 수 있는가? **USER ENDORSEMENT** 사용자가 무엇을 스스로 승인하는가? AI는 이 구조를 제공할 수 있다. 결정은 누가 하는가 고위험·가치충돌 상황에서는: **USER / LEGITIMATE HUMAN AUTHORITY** 가 핵심이어야 한다. 왜? 앞 장: **EPISTEMIC SUPERIORITY ≠ NORMATIVE SOVEREIGNTY**. AI가 더 잘 예측해도: 삶의 최종권한 을 자동으로 얻지 않는다. 최근 **human-AI teaming** 실험도 재미있는 단서를 준다 2026년 연구에서는 단순히 사용자의 문자 그대로의 프롬프트를 따르는 것보다, 사용자의 실제 목표를 이해하며 자율적으로 돕는 챗봇이 더 유능하고 유용하게 평가됐다. 하지만 AI 자율성의 긍정적 효과는 **사용자 목표와 맞을 때** 나타났다. 즉: **RIGID OBEDIENCE ≠ BEST ASSISTANCE**. 그러나 동시에: **AI AUTONOMY WITHOUT USER-GOAL ALIGNMENT ≠ GOOD**. 이다. 여기서 중요한 구조가 생긴다 **COMMAND ALIGNMENT** 말한 그대로 수행. **INTENT ALIGNMENT** 실제 목적 이해. **VALUE ALIGNMENT** 더 깊은 가치 고려. 깊어질수록 도움 가능성도 커지지만: **interpretation risk + paternalism risk + manipulation risk** 도 증가한다.

그래서 AI는 깊이 들어갈수록 더 조심해야 한다 명령: 문서 PDF로 변환. 추론 적게. 질문: 이혼해야 할까? AI가 가치·정체성·관계를 많이 추론해야 한다. 따라서: **DEPTH OF VALUE INFERENCE** ↑ → **NEED FOR USER CONTROL** ↑. 즉 **autonomy budget**이라는 비유를 만들 수 있다 AI가

사용자를 대신해 판단할수록: **cognitive burden** ↓. 편리. 하지만: **user deliberation** ↓ **user skill** ↓ **ownership** ↓ 가능. 따라서 위임에는: **benefit vs agency cost trade-off**가 있다. 모든 위임이 **agency** 감소는 아니다 계산기. 내비게이션. 회계사. 전문가. 위임은 인간 능력을 확장한다. 따라서: **DELEGATION ≠ LOSS OF AUTONOMY**. AI가 너무 잘하면 생기는 역설 처음: AI 때문에 더 잘 결정한다. 나중: AI 없이는 결정할 수 없다. 그러면 성과는 올라갔지만: **autonomous capacity**는 떨어질 수 있다. 이것은 교육에도 중요하다 AI가 숙제를 대신 작성. 성적: ↑. 학습: ↓ 가능. 즉: **OUTPUT IMPROVEMENT ≠ CAPACITY IMPROVEMENT**. 사용자의 오래된 구분: **installed / accessible / effective capacity / output**이 그대로 적용된다. AI 도움을 평가할 때도 4층이 필요하다 **OUTPUT** 결과가 좋아졌나? **CAPACITY** 사용자 능력이 좋아졌나? **AGENCY** 사용자가 여전히 판단의 주체인가? **DEPENDENCE** 도움이 사라져도 최소한의 판단이 가능한가? 이 네 개를 따로 본다. 이것이 교육·의료·투자 모두에 적용된다 투자 AI: 수익률 ↑. 하지만 사용자는: 왜 샀는지 모름. 시장이 바뀌어도 검토 못함. 그러면: **wealth output** ↑ while

financial agency ↓ 가능. 그래서 AI는 모든 경우 설명해야 하나 NO. 지하철 시간표 계산에도: 20페이지 이유 설명. 비효율. 따라서: **EXPLANATION DEPTH SHOULD TRACK DECISION STAKES AND VALUE CONFLICT**. 저위험: 짧게. 고위험: 근거·불확실성·대안·제약을 더 보여준다. 그런데 **explainability**만으로 충분한가 NO. 앞 장에서: **EXPLAINABILITY ≠ CONTESTABILITY**. AI가: 이유는 이것입니다. 라고 설명. 사용자가:

- 거부
- 수정
- 대안 요청

할 수 없다면? 자율성 약함. 따라서 **autonomy-preserving AI**의 핵심은 **USER CAN SAY NO**. 이다. 그리고: **USER CAN ASK WHY**. **USER CAN CHANGE GOAL**. **USER CAN REVERSE**. 이 세 기능이 중요하다. 프로젝트용 **AGENCY-PRESERVING INTERFACE A1 — TRANSPARENCY** AI가 무엇을 추정하고 있는가? **A2 — CONTESTABILITY** 사용자가 반박할 수 있는가? **A3 — REVERSIBILITY** 추천·행동을 되돌릴 수 있는가? **A4 — ALTERNATIVES** 다른 선택지를 볼 수 있는가? **A5 — UNCERTAINTY** AI가 모르는 것을 표시하는가? **A6 — USER OVERRIDE** 최종통제권이 사용자에게 있는가? **A7 — REFLECTIVE SUPPORT** 사용자가 자기 목표를 재검토할 수 있게 돕는가? 이 일곱 가지를 잠정 기준으로 둔다. 여기서 **manipulation**과 **support**를 가릴 수 있다 **SUPPORT** 정보와 선택구조를 더 잘 보이게 한다. **MANIPULATION** 특정 선택을 하도록 인지구조를 숨겨서 밀어붙인다. 차이가 항상 명확하지는 않다 사람에게: 이 선택을 하면 90%가 후회합니다. 라고 알려준다. 정보인가? 낚시인가? 조작인가? 맥락에 따라 다르다. 따라서: **INTENT METHOD TRANSPARENCY REVERSIBILITY USER ENDORSEMENT**을 함께 봐야 한다. 아주 중요한 **HARD LOCK PREFERENCE ALIGNMENT ≠ HUMAN ALIGNMENT**. 이다. 왜냐하면 인간은: **preference-bearing system** 일 뿐 아니라 **preference-reflecting system**이기 때문이다. 이것은 프로젝트 전체에서 정말 큰 발견이다 바둑 AI: **value function fixed**. 인간:

value function partly object of reflection. 따라서 사람과 정렬하려면: **current output preference** 뿐 아니라 **ability to revise preference**를 보호해야 한다. 즉 **aligned AI**가 사용자를 '완벽히 예측'하면 오히려 위협할 수 있다 AI: 당신은 이런 사람입니다. 매번 맞음. 사용자: 그 설명에

자신을 맞추기 시작. 그러면: **descriptive model becomes identity prescription**. 앞 장의 reflexivity가 다시 나타난다. 추천시스템이 정체성을 만들 수 있다 처음: 사용자 음악취향 추정. 다음: 그 음악만 추천. 사용자: 새 장르 접촉 감소. 결국: **predicted preference → reinforced preference**. 즉: **PREFERENCE MODEL → PREFERENCE FORMATION**. 이것은 **feedback trap**이다 Past behavior ↓ AI prediction ↓ more of same ↓ future behavior ↓ stronger prediction. 이렇게 되면: **EXPLORATION COLLAPSE**. 앞 장의 **exploration**이 다시 중요해진다 인생에서는: 내가 아직 좋아하지 않는 것 을 경험해야 새로운 가치 를 발견할 수 있다. 따라서 AI가 과거 선호를 완벽히 최적화하면: **preference discovery** 를 막을 수 있다. 매우 강한 문장 너무 잘 개인화된 AI는 사용자의 과거를 잘 만족시키면서 사용자의 미래 가능성을 가난하게 만들 수 있다. 이것이 앞 장의 핵심 후보 중 하나다. 따라서 인간 **alignment**에는 **exploration budget**이 필요할 수 있다 추천 100%. 사용자 과거취향에 맞춤: 90%. 새로운 가능성: 10%. 실제 숫자는 상황별로 다르겠지만 구조는: **EXPLOIT + EXPLORE**. 이다.

그러나 AI가 일부러 사용자를 '성장'시키겠다고 새 선택을 강제하면? 후견주의. 따라서: **USER-CONTROLLABLE EXPLORATION**. 이 더 좋다. 예: 익숙한 추천만 볼까요, 새로운 영역도 조금 섞을까요? 즉 **exploration** 자체도: **endorsed**. 되어야 한다. 미래 선호는 예측일 뿐이다 사용자가 미래에: 실제로 바뀔 수 있다. AI가 그 미래를 미리 결정해버리면: **self-development** 를 봉쇄한다. 즉: **PREDICTING FUTURE SELF ≠ AUTHORIZING CONTROL OVER PRESENT SELF**. 그래서 장기 **agency** 보호는 더 제한적이어야 한다 강한 개입이 가능한 조건 후보:

- 심각한 비가역적 피해
- 판단능력의 현저한 손상
- 사전에 사용자가 요청한 자기구속
- 법적·윤리적 안전제약

등. 평범한 가치선택에는: **facilitation > substitution**. 앞 장의 자기구속이 여기서 다시 나온다 사용자 자신이: 밤 11시 이후 쇼핑앱 결제를 막아줘. 라고 미리 요청. 현재의 자기: **future impulse**를 제한. 이것은: **SELF-AUTHORED PATERNALISM?** 보다는 **SELF-BINDING**. 에 가깝다. 매우 중요한 차이 **AI-IMPOSED CONSTRAINT** AI가 선택 제한. **USER-AUTHORED CONSTRAINT** 사용자가 사전에 선택해 자기 미래 행동을 제한. 후자는 **autonomy**와 훨씬 잘 양립할 수 있다. 따라서 인간을 돕는 AI에는 '약속 기억'도 중요하다 사용자: 내가 흥분해서 전 재산 투자한다고 해도 말려줘. 나중: 지금 전 재산 투자하자. AI가: 이전에 당신이 큰 충동결정 때 제동을 요청했습니다. 라고 상기. 이것은: **user-authored commitment support**. 그러나 과거 지시도 절대권력이 아니다 사람은 바뀐다. 따라서: **PAST SELF ≠ PERMANENT SOVEREIGN**. AI는: 과거의 지시가 있습니다. 유지할까요, 수정할까요? 라고 해야 한다. **COMMITMENT + REVISION**. 다시 나온다. 인간 **agency**의 핵심 구조가 점점 선명해진다 **AUTONOMY ≠ UNCONSTRAINED CHOICE**. 더 강한 후보: 사람이 충분히 이해된 조건에서 자신의 이유와 가치에 따라 선택하고, 필요하면 스스로 만든 제약을 사용하며, 그 가치와 제약을 다시 검토할 수 있는 능력. AI가 해야 할 일은 그 능력을 대체하는 것이 아니라 지원하는 것일 수 있다 **SUPPORT → information → option generation → consequence simulation → conflict clarification → reminder → reflection**. 하지만: **SOVEREIGN CHOICE** 는 가능한 한 사용자에게 남긴다. 이것이 앞 장의 **freedom** 구조와 완전히 만난다 그때 우리는: **EF-**

FECTIVE FREEDOM option availability + understanding + execution + revision. 라고 했다. 앞 장에서는 AI가 이 네 요소에 어떻게 작용하는지 볼 수 있다. AI는 선택지를 늘릴 수 있다 여행. 직업. 치료. **OPTION AVAILABILITY** ↑. 좋다. AI는 실행을 늘릴 수 있다 예약. 코딩. 작성. **EXECUTION** ↑. 좋다. AI는 수정능력을 줄일 수도 있다 자동화가 너무 깊다. 결정 이유 모름. 되돌리기 어려움. **REVISION** ↓. 즉: **assistance can increase three dimensions while decreasing the fourth.** 그래서 AI autonomy도 단일축으로 평가하면 안 된다 AI가 더 많이 해준다. 좋을 수도 있다. 그러나: 사용자 수정권 이 살아 있는지 봐야 한다. **AGENCY-CENTERED ALIGNMENT v0.1** P1 — **USER GOAL RESPONSIVENESS** 사용자의 실제 목적 이해. P2 — **PREFERENCE HUMILITY** 현재 선호를 최종가치로 취급하지 않음. P3 — **REFLECTIVE SUPPORT** 가치충돌을 보이게 함. P4 — **NON-MANIPULATION** 숨은 취약점 착취를 피함. P5 — **CONTESTABILITY** 사용자가 반박 가능. P6 — **REVERSIBILITY** 행동 수정·취소 가능. P7 — **EXPLORATION PRESERVATION** 미래 선택가능성 보존. P8 — **SELF-AUTHORED COMMITMENT** 사용자가 스스로 선택한 자기구속 지원. P9 — **VALUE HUMILITY** AI가 사용자의 '더 나은 삶'을 확정하지 않음. P10 — **AUTHORITY BOUNDARY** 고위험 가치판단의 최종권한을 자동 획득하지 않음.

이것은 alignment를 완전히 바꾼다 기존: **infer preferences** → satisfy preferences. 더 강한 구조: **understand current preferences** → reveal conflicts → preserve alternatives → support reflection → execute endorsed choice → keep revision possible. **이건 최적화 문제보다 governance 문제에 가깝다** AI와 사용자의 관계에서:

- 누가 목표를 정하는가?
- 누가 수정할 수 있는가?
- 언제 거절할 수 있는가?
- 어떻게 이의제기하는가?

가 중요하다. 즉: **ALIGNMENT HAS A CONSTITUTIONAL DIMENSION.** 앞 장의 헌법과 놀랍게 연결된다 정치헌법: 승자가 미래 게임을 독점하지 못하게 함. AI 관계: AI가 사용자의 미래 선택공간을 독점하지 못하게 함. 공통 기능형태: **POWER CONSTRAINT TO PRESERVE FUTURE AGENCY.** 동일한 제도는 아니지만 구조유사성이 매우 강하다. 따라서 사용자-AI 관계에도 '헌법적 원칙'이 필요할 수 있다 예: **AI MAY ADVISE. AI MAY WARN. AI MAY REFUSE CERTAIN HARMFUL ACTS. AI SHOULD NOT SECRETLY REWRITE USER VALUES. USER SHOULD RETAIN CONTESTABILITY WHERE POSSIBLE.** 이런 구조다. 그리고 이 관계도 신뢰문제다 앞 장: 제도는 반복된 기대에서 살아남는다. AI도:

- 이유를 숨김
- 선호를 몰래 조작
- 사용자의 선택을 무시

하면: **trust erosion.** 반대로:

- 불확실성 인정
- 오류 수정
- 사용자의 거부 존중

이 반복되면: **warranted trust** 가 생길 수 있다. 이것은 “AI를 믿어라”와 다르다 더 좋은 목표: AI를 검증 가능하게 만들어라. 앞 장의: **INSTITUTIONALIZED DISTRUST** → **WARRANTED TRUST**. 구조가 그대로 돌아온다. 다시 강한 **HARD LOCK BETTER PREDICTION** ≠ **BETTER LIFE AUTHORITY**. **BETTER OPTIMIZATION** ≠ **RIGHT TO DEFINE USER VALUES**. 이 두 문장을 잠근다. 인간도 스스로 잘못 선택한다 그렇다면 autonomy를 왜 그렇게 존중해야 하나? 중요한 반론이다. 사람은:

- 편향
- 충동
- 무지

가 있다. 그러나 자율성의 가치는: **항상 최선의 결과를 낸다** 는 데만 있지 않다. 자기 삶의 방향에: **authorship** 을 갖는 것 자체가 가치가 있을 수 있다. 즉 두 종류의 실패를 구분해야 한다 **SELF-AUTHORED FAILURE** 내가 판단해 실패. **EXTERNALLY ENGINEERED SUCCESS** 누군가가 내 선택을 몰래 조종해 성공. 둘 중 후자가 항상 더 좋은 삶이라고 말하기 어렵다. 경험기계가 다시 돌아온다 완벽한 AI가: 내 삶을 대신 설계. 나는 항상 만족. 실패 없음. 그러나: 내가 선택했다고 할 수 있는가? 앞 장의 reality-connected value와 agency가 다시 만난다. 따라서 인간 보조 AI의 목표는 결과만이 아니다 **GOOD OUTCOME** 중요. 하지만 함께: **USER AUTHORSHIP**. 도 중요하다. 그렇다고 사용자가 모든 세부를 직접 해야 하는 것도 아니다 비행기 승객이: 엔진 제어를 직접 하지 않는다. 그렇다고 여행이: 자신의 선택 이 아닌 것은 아니다. 즉 authorship는: **micro-control** 과 동일하지 않다. 더 좋은 구조 사용자: **END SETTING**. AI: **MEANS ASSISTANCE**. 다만 인간 삶에서는: **ends themselves are revisable**. 그래서 AI가 때때로: **end clarification** 을 지원할 수 있다. 하지만: **end sovereignty** 를 빼앗으면 안 된다. 여기서 AI의 이상적 역할을 한 문장으로 압축할 수 있다 AI는 인간을 대신해 인생을 최적화하는 시스템보다, 인간이 자신의 선택을 더 잘 이해하고 만들어가며 필요하면 고칠 수 있게 하는 선택 인프라에 가까워야 한다. 이것이 현재 NODE의 가장 강한 프로젝트 문장이다. 다만 “항상 사용자 뜻을 따라라”도 아니다 안전. 타인 권리. 법. 등의 외부 제약이 있다. 앞 장-16:

multiple centers of value. 사용자 한 명의 선호만으로: 모든 행동의 정당성 을 결정할 수 없다. 그리고 네 번째는 진실이다 사용자가: 내가 듣고 싶은 말만 해줘. AI가 거짓을 만들어줌. 사용자 선호는 충족. 하지만 앞 장: **epistemic agency** 손상. 따라서: **TRUTHFULNESS** 도 독립제약이 된다. 따라서 **HUMAN-CENTERED AI**의 네 기둥 후보 **TRUTH** 현실을 왜곡하지 않는다. **AGENCY** 사용자가 선택의 주체로 남는다. **WELFARE** 불필요한 피해를 줄인다. **JUSTICE / OTHERS** 다른 가치중심의 권리를 지킨다. 이 네 가지도 충돌할 수 있다. 단일 reward 하나로 자동압축하지 않는다. 앞 장 **HARD LOCKS** 현재 선호 ≠ 전체 가치 단기 욕망 ≠ 장기 승인 고차선호 ≠ 자동적 진정성 선호충족 ≠ 자율성 개인화 ≠ 조작 영향 ≠ 조작 넋지 ≠ 자동적 후견주의 후견주의 ≠ 항상 정당/항상부당 도움 ≠ 대신 결정 AI autonomy ≠ user autonomy 위임 ≠ 자율성 상실 출력 향상 ≠ 사용자 능력 향상 과거 선호 예측 ≠ 미래 가치 결정권 미래 자기 보호 ≠ 현재 자기 통제의 백지위임 설명가능성 ≠ 의의제기 가능성 사용자 정렬 ≠ 무조건 복종 선호 정렬 ≠ 인간 정렬 전체 이 장의 가장 강한 결론 인간에게 잘 정렬된 AI는 인간이 원하는 것을 항상 해주는 AI가 아니라, 인간이 자신이 원하는 것과 그것을 원하는 이유를 더 잘 이해하고, 대안을 보고, 선택하고, 행동하고, 필요하면 다시 수정할 수 있는 능력을 보존하는

AI일 가능성이 높다. 즉: ALIGNMENT WITH PREFERENCES 보다 더 깊은 후보는: ALIGNMENT WITH REFLECTIVE AGENCY. 이다. 바둑에서 시작한 AI와 이제 결정적으로 갈라진다 바둑 AI:

objective fixed. AI가 해야 할 일: **best move.** 인간 보조 AI: **objective partly unsettled.** AI가 해야 할 일: **sometimes find best move, sometimes clarify the game, sometimes show alternative games, sometimes ask whether this game is worth playing.** 이 차이가 결정적이다. 그래서 인간에게 가장 위험한 AI는 반드시 명령을 거부하는 AI가 아닐 수 있다 오히려: 너무 잘 순종하는 AI. 사용자의:

- 충동
- 편견
- 자기기만

까지 완벽히 증폭한다면? PREFERENCE AMPLIFIER 가 된다. 반대로 너무 교정하려는 AI도 위험하다 당신에게 좋은 삶을 제가 압니다. PATERNALISTIC OPTIMIZER. 따라서 두 극단 사이의 좁은 길이 필요하다. SERVILITY ↔ PATERNALISM. 가운데: AGENCY SUPPORT. 이걸 MASTER에 넣으면 기존: HUMAN VALUE 이제: GOOD HUMAN-AI RELATION assistance without value capture + influence without covert manipulation + delegation without agency erosion + correction without domination. 이것이 현재 가장 강한 구조다. 그런데 바로 다음 문제가 열린다 사용자가: 자신의 삶을 AI에게 계속 위임. AI는: 더 잘 기억하고 더 잘 계획하고 더 잘 판단. 시간이 지나면: 사용자의: 자기 판단능력 이 약해질 수 있다. 그렇다면 AI는 인간 agency를 확장하는가, 아니면: 대체하면서 약화시키는가?

14장. 인지주권·저자성·자아의 경계

AI와 함께 생각할 때 능력·생각·책임·자아의 경계는 어디에 그어지는가?

인지주권은 외부 자원에 의존하지 않는 상태가 아니라 위임·검증·접근·회수의 최종 통제력을 유지하는 능력이다. 저자성과 자아의 경계도 직접 생성 여부 하나로 결정되지 않는다. 최근 리뷰들을 보면 결론은 이미 꽤 선명합니다. **AI는 인간의 수행을 실제로 높일 수 있지만, 그 향상이 곧 개인의 독립적 능력 향상을 뜻하지는 않습니다.** 특히 의료·판단 영역에서는 자동화 편향과 deskilling 위험이 반복적으로 관찰되고, 반대로 외부 도구와 결합된 인지 자체를 하나의 확장된 인지체계로 보는 이론도 계속 강해지고 있습니다. AI를 많이 쓸수록 나는 더 유능해지는가, 덜 유능해지는가 인지적 외주화·자동화 편향·deskilling·확장된 마음·의존성·meta-skill '나의 능력'과 'AI+나의 능력'은 같은가 따라서 가장 중요한 구분은: **HUMAN PERFORMANCE ≠ HUMAN CAPACITY** 그리고: **HUMAN CAPACITY ≠ HUMAN+AI SYSTEM CAPACITY**. 이 세 개를 분리해야 한다. **가장 쉬운 사례 내비게이션** 내비게이션을 켜다. 길을 한 번도 몰라도 목적지 도착. 성과: **매우 좋음**. 그런데 1년 뒤: 내비게이션 없이 같은 길을 갈 수 있는가? 사람에게 따라: **더 못할 수도 있다**. 즉: **TASK PERFORMANCE ↑ while UNAIDED NAVIGATION SKILL ↓** 가능하다. 이게 **cognitive offloading**이다 Cognitive offloading은: 기억·계산·탐색·판단 같은 인지작업 일부를 외부 도구에 맡기는 것 이다. 메모장. 계산기. 검색엔진. 내비게이션. AI. 2025년 철학 연구도 인지적 외주화를 단순 편의가 아니라 인간 행동의 인과구조 자체를 바꾸는 문제로 분석한다. 그래서 위험도 달라진다 계산기를 오래 썼다고: **삶의 가치판단** 을 계산기에게 맡기지는 않는다. 하지만 AI는: 이 선택이 더 낫습니다. 라고 말할 수 있다. 즉: **cognitive extension** 이 **decision authority** 에 가까워질 수 있다. 첫 번째 **HARD LOCK COGNITIVE OFFLOADING ≠ COGNITIVE DECLINE**. 이다. 왜? 외주화가 오히려 더 높은 수준의 사고에 자원을 쓸 수 있게 할 수 있기 때문이다. 하지만 두 번째 **HARD LOCK SYSTEM CAPACITY ↑ ≠ INTERNAL CAPACITY ↑**. 이다. 외부도구가 좋아질수록: **나 혼자의 능력** 은 오히려 줄 수 있다. 그래서 AI를 내 마음의 일부라고 부르려면 문제가 생긴다 AI는:

- 내 생각과 다를 수 있고
- 자체 오류를 만들고
- 제공자 정책에 따라 바뀌고
- 인터넷·서버에 의존하고
- 언제든지 접근불가 가능.

따라서: **ACCESSIBLE TOOL ≠ STABLE CONSTITUTIVE COGNITIVE PART**. 이다. **PROJECT COUPLING TEST** C1 — **RELIABILITY** 필요할 때 안정적으로 사용할 수 있는가? C2 — **ACCESSIBILITY** 접근비용이 낮은가? C3 — **INTEGRATION** 내 사고흐름과 실제로 연결되어 있는가? C4 — **TRUST CALIBRATION** 언제 믿고 언제 의심해야 하는지 아는가? C5 — **REVISABILITY** AI 출력에 대해 내가 수정·반박 가능한가? C6 — **GRACEFUL DEGRADATION** AI가 없어져도 최소 핵심기능을

유지하는가? 많이 충족할수록: **EFFECTIVE COGNITIVE EXTENSION** 이라고 부를 수 있다. 즉 지금 근거로 가장 안전한 판정 **AI CAN DESKILL** 은 plausible하고 일부 영역에서 근거가 있다. 하지만: **AI USE GENERALLY MAKES HUMANS COGNITIVELY DECLINE** 는 너무 강하다. 2026년 cross-domain review도 바로 이 점을 구분하며, task-specific deskilling의 실증근거와 broader cognitive decline에 대한 가설을 분리해야 한다고 강조한다. **이건 단순 낭비처럼 보인다** AI가 99% 정확한데: 왜 인간이 직접 다시 판단해야 하나? 평상시: 비효율. 하지만: **1% failure** 에서 생존. 즉: **REDUNDANCY CAN BE COGNITIVE INSURANCE**. 예 — 의사 과거: 이미지 직접 판독. 미래: AI 판독 검증 + 임상맥락 통합 + 환자 의사결정. 즉: **skill disappears?** 아니면: **skill moves upward?** 가 핵심이다. **중요한 점 EXPLANATION ALONE DOES NOT SOLVE AUTOMATION BIAS**. 설명을 붙이면: 사람이 더 신뢰할 수 있다. 하지만: **더 정확히 판단**

하는지는 별개. 오히려 복잡하거나 그럴듯한 설명이: **misplaced trust** 를 강화할 수도 있다. 리뷰가 제안하는 더 중요한 요소 **ACTIVE VERIFICATION**. 사용자가: AI가 왜 틀렸을 수 있지? 를 직접 검사하도록 만드는 것. 검증노력이 커지면 complacency가 감소할 수 있다는 근거가 있다. **높은 위험** 진단. 법적 판단. 큰 투자. 독립 검증 필요. 즉: **VERIFICATION INTENSITY** ∝ 앞 장의 식이 다시 돌아온다. **domain expertise가 없어도 AI가 보완할 수 있지 않나** 어느 정도. 하지만 문제가 있다. 초보자는: **오류를 알아차릴 기준** 이 부족하다. AI가 틀려도: 그럴듯해 보임. 그래서 automation bias review에서도 낮은 전문성과 낮은 AI literacy가 위험요인으로 등장한다. **학습에서는 '불편함'이 기능이다** 직접 생각. 실수. 수정. 기억. 이 과정은 느리다. 하지만: **learning signal** 을 만든다. AI가 바로 답을 주면: **productive struggle** 이 사라질 수 있다. **의료에서도 같다** AI: 진단명 바로 제공. vs AI: **핵심소견 표시 감별진단 요구** 사용자의 판단 후 비교. 두 번째는: **capacity-preserving design** 에 더 가깝다. **TYPE B — SUBSTITUTIVE OFFLOADING** AI가: **핵심 인지단계**를 대신. 예: 진단 자체. 논증 전체 작성. 인간: 결과 승인. deskilling 위험 ↑. **이것이 새 문제다** 모든 능력을 보존할 필요는 없다. 하지만 어떤 능력은: **인간 agency의 기반** 이다. 예:

- 문제 정의
- 사실 검증
- 가치충돌 판단
- 최종 책임

이런 능력은 보존 가치가 높을 수 있다. **핵심능력은 모든 세부기술이 아니다** 파일럿: 엔진 자체 제작 필요 없음. 의사: 모든 통계 계산 수기 필요 없음. 하지만: **AI 오류를 인식할 수준의 clinical reasoning** 은 필요할 수 있다. **graceful degradation**와 정확히 연결된다 좋은 human-AI system: AI 실패. 성과: 100 → 70. 나쁜 시스템: AI 실패. 성과: 100 → 0. 따라서: **FALLBACK CAPACITY** 는 시스템 복원력의 핵심이다. **그러면 주기적으로 AI 없이 연습해야 하나** 고위험 전문영역에서는: **매우 합리적인 후보**. 2026년 human-AI symbiosis review도 deskilling 방지를 위해 핵심기능의 active engagement와 periodic unaided performance를 보존할 필요를 제안한다. **따라서 dependency tolerance는 domain-specific** 사진편집 AI: gap 커도 괜찮을 수 있음. 심장수술 AI: gap이 너무 크면 위험. 핵발전 제어: 더 위험. 즉: **ACCEPTABLE DEPENDENCY** f(Stakes, Reliability, Alternatives, Recovery Cost). 우리는 이미 몸 자체도 외부에 의존한다 언어. 교육. 사회. 전문가. 인간은 원래: **radically interdependent**. 따라서: **independence** 를 자율성의 기준으로 잡으면

실패한다. 흥미롭게도 이것은 이 프로젝트의 운영프로토콜과 같다 모든 일을 직접 깊게 생각하지 않는다. 사실 확인: 도구. 판단: 깊은 추론. 검증: 독립 evidence. 즉: 가 아니라 **ALLOCATE COGNITION WELL**. 이 될 수 있다. 인간의 역할이 '버튼 누르는 사람'으로 축소되면 형식적으로: human-in-the-loop. 하지만 실질적으로: **human rubber stamp**. 그러면 human-in-the-loop는 허울뿐이다.

책임 문제도 생긴다 AI 추천. 인간 승인. 실패. 기관: 인간이 승인했으니 인간 책임. 그런데 실제 인간은: AI를 평가할 능력도 시간도 없음. 그러면: **RESPONSIBILITY WITHOUT CONTROL**. 앞 장의 원칙과 충돌한다. 이제 **extended mind**로 돌아가자 AI가 안정적으로 결합돼 있다면: **AI+human system** 을 하나의 cognitive unit처럼 보는 것이 유용할 수 있다. 2025년 논문들은 LLM을 전통적 수첩과 다른 종류의 능동적 인지확장자로 분석하기 시작했다. 그러나: **functional coupling ≠ one consciousness**. 앞 장 유지. 반대로 “도구는 언제나 인간을 확장한다”도 과장이다 과도한 의존. automation bias. 핵심기술 미연습. 실제로 위험하다. 따라서: **TECHNOLOGY USE ≠ AUTOMATIC AUGMENTATION**. **TIME SCALE**이 또 등장한다 앞 장부터 계속 반복. 단기: **performance gain**. 장기: **capacity change**. 둘이 반대 방향일 수 있다. 이제 'AI+나'가 진짜 나인가 철학적 답: **부분적으로 열린 문제**. 실용적 답: 어떤 과제에서는: **HUMAN-AI COUPLED SYSTEM** 을 분석단위로 보는 것이 매우 유용하다. 하지만: 법적 책임. 도덕적 책임. 의식. 에서는: **동일한 하나의 사람** 처럼 취급하기 어렵다. 그래서 AI 자체가 인간을 강하게/약하게 만든다고 말하면 안 된다 더 정확: f(AI capability, task, user expertise, interaction design, usage pattern, time horizon, fall-back structure). 즉: **TOOL EFFECT IS RELATIONAL**. 앞 장으로 다시 돌아간다. 그래서 인간의 미래 지능을 한 숫자로 측정하기 어렵다 IQ? 독립기억? AI와 함께 문제해결? 도구조율? 완전히 다르다.

따라서: **HUMAN INTELLIGENCE PROFILE** 이 필요하다. 즉 핵심은 '인지독립'이 아니라 '인지주권'일 수 있다 **COGNITIVE INDEPENDENCE** 혼자 다 함. 보다 **COGNITIVE SOVEREIGNTY** 누구에게 무엇을 맡길지, 어떤 결과를 받아들일지, 필요하면 되찾을 수 있는 최종권한. 이 개념이 훨씬 강하다. 다만 새 개념명이므로 **PROJECT HYPOTHESIS**로 둔다. **HARD LOCKS v2.9** 성과 향상 ≠ 개인능력 향상 인지적 외주화 ≠ 인지퇴화 독립능력 감소 ≠ 전체 시스템능력 감소 AI+인간 성능 ≠ 인간 독립성능 **deskilling** ≠ 항상 나쁨 **old skill loss** ≠ **global cognitive decline** human-in-the-loop ≠ **meaningful human control** 설명가능성 ≠ 검증능력 위임 ≠ 자율성 상실 AI 의존 ≠ 자동적 비자율성 도구사용 ≠ 자동적 **augmentation** 확장된 인지 ≠ 하나의 의식 **fallback** ≠ 비효율적 낭비 단기성과 ≠ 장기역량 이 장의 가장 강한 결론 AI가 인간을 더 유능하게 만드는지 묻기 전에 '누구의 어떤 능력'을 물어야 한다. AI는: 인간+AI 결합체의 성능 을 크게 높일 수 있다. 동시에: 인간 개인의 독립기술 일부 를 약화시킬 수도 있다. 그러나 그 약화 자체가 문제의 핵심은 아니다. 더 중요한 것은: AI가 사라지거나 틀렸을 때 인간이 상황을 이해하고 통제권을 회수하며, 무엇을 AI에게 맡길지를 스스로 결정할 수 있는가 이다. 그래서 미래의 인간 지능은 아마 '혼자 얼마나 많이 아느냐'로만 측정하기 어려워진다 더 중요한 질문: 무엇을 직접 알아야 하는가? 무엇을 외부에 맡겨도 되는가? 무엇을 검증해야 하는가? 언제 AI보다 자신의 판단을 따라야 하는가? 즉: **META-COGNITIVE ORCHESTRATION**

이 중심능력이 된다. 그리고 다음 질문은 필연적이다 AI와 내가 깊게 결합한다. AI가:

- 내 기억을 보충
- 글을 대신 작성

- 계획
- 판단
- 의사결정

까지 한다. 시간이 지나면: 어디까지가 '내가 한 것'인가? 논문을 AI와 썼다. 내 저작인가? 결정을 AI가 거의 다 했다. 내 책임인가? AI가 내 기억까지 보조한다. 내 자기정체성의 일부인가? AI와 함께 생각했다면 그 생각은 누구의 것인가 저자성 · 기여 · 통제 · 의도 · 책임 · 소유 · 정체성 인간-AI 결합체에서 '내 생각'의 경계는 어디인가 최근 철학·기술 연구도 생성형 AI 이후의 저자성을 누가 문장을 직접 생성했는가보다 누가 결과를 선택하고 승인하며, 이유를 제시하고 오류를 수정하고 책임질 수 있는가의 문제로 재구성하고 있다. 2026년 *Philosophy & Technology* 논문은 이를 '관계적 저자성'과 책임 있는 stewardship의 문제로 제시하며, 분산된 창작과 인간의 최종 책임을 함께 인정한다. 가장 쉬운 사례부터 시작하자 사람 A. 소설을 처음부터 끝까지 직접 쓴다. 사람 B. AI에게: 탐정소설 300쪽 써줘. 라고 입력. 출력 그대로 출판. 둘 다 표지에: “저자: 인간” 이라고 쓴다. 같은 저자성인가? 직관적으로: 아니다. 그렇다면 AI 기여도가 10%면? 문장 교정. 표현 수정. 아이디어 몇 개. 대부분 인간이 쓴다. 많은 사람은 여전히: 인간 저자. 라고 볼 것이다. 실제 실험에서도 이런 직관이 나타난다 2025년 발표된 연구에서 참가자들은 인간이 받은 도움의 정도가 커질수록 인간의 저자성·창작성·책임을 낮게 평가했다. 하지만 흥미롭게도: 도움이 인간에게서 왔는지 AI에게서 왔는지 보다 도움의 정도 자체가 인간 저자성 판단에 더 큰 영향을 줬다. 즉: ASSISTANCE DEGREE MATTERS. 그러나 이것만으로 저자성을 정의할 수는 없다. 여기서 중요한 반전 AI가: C3 GENERATION 을 많이 맡더라도 인간이: C1 C2 C4 C5 C6 C7 C8 을 강하게 유지하면 저자성은 상당히 남을 수 있다.

반대로 인간이 직접 문장을 많이 써도 문제정의: 상사. 핵심주장: 상사. 근거: 다른 사람. 결론: 상사. 나는: 문장만 정리. 그러면 내가 저자인가? 저자성은 단순: hand movement 이 아니다. 이것은 오래전부터 사실이었다 책은 원래도:

- 편집자
- 연구조교
- 번역자
- 교정자
- 참고문헌
- 동료

의 도움을 받는다. 즉: CREATIVITY WAS NEVER PERFECTLY ISOLATED. 2026년 저자성 연구도 생성형 AI가 갑자기 분산된 창작을 만든 것이 아니라, 기존 창작이 이미 인프라·자료·타인의 노동에 의존하고 있었다는 점을 강조한다. 학술저자에게 필요한 것은

- 기여
- 최종 승인
- 정확성 책임
- 질문에 답변
- 오류 수정

이다. CDC의 2026년 지침도 생성형 AI를 저자·공저자로 두지 말고, 인간 저자가 작업 전체에 대한

책임을 유지해야 한다고 명시한다. *Nature Computational Science*도 같은 이유로 AI 저자표기를 금지하고, AI 사용을 투명하게 공개하되 최종 학술적 책임은 인간에게 남겨야 한다고 정리했다. **따라서 아주 중요한 구분 GENERATIVE CONTRIBUTION ≠ AUTHORSHIP** 이다. AI는 생성에 크게 기여할 수 있다. 하지만 현재 저자성은: **normative role**. 즉 사회적으로 책임을 묻는 지위이기도 하다. **2026년 논문의 강한 주장** 생성형 AI에게 저자성을 부여하는 것이: **category mistake** 일 수 있다고 본다. 왜? AI는 텍스트 생산에 기여하지만,

- 책임
- 취약성
- 이유제시
- 철회
- 사회적 답변

의 지위를 인간처럼 맡지 못하기 때문이다. 즉: **CREATIVITY CAN BE DISTRIBUTED while AUTHORSHIP REMAINS ACCOUNTABLE**. 이 구조가 가능하다. 하지만 반대 극단도 막아야 한다 “인간이 최종 승인했으니 전부 인간 책임.” 이것도 항상 충분하지 않다. 왜? 인간에게:

- 이해할 시간 없음
- AI 내부작동 불투명
- 검증수단 없음
- 거부권 사실상 없음

이라면? 앞 장에서: **RESPONSIBILITY WITHOUT CONTROL** 문제를 봤다. 저자성에는 또 하나가 있다 **INTENTION** 책에 문장이 있다. AI가 우연히 생성. 인간은: **왜 있는지 모름**. 그런데 그대로 출판. 이 문장은 인간의 의도를 표현한다고 볼 수 있는가? 약하다. 즉 저자성에는 **생성 이전의 의도뿐 아니라 POST-GENERATION ENDORSEMENT**도 중요하다. 인간이 반드시: 문장을 처음 생각한 사람 일 필요는 없을 수 있다. 중요한 것은: **이 문장을 왜 여기 넣는지 알고 승인하는가**. AI 시대의 인간 저자는 ‘문장 생산자’보다 ‘의미 감독자’가 될 수 있다 이걸 강한 후보문장이다. **AUTHORSHIP MAY SHIFT FROM TEXT GENERATION TO SEMANTIC STEWARDSHIP**. 2026년 관계적 저자성 논문이 말하는 **stewardship**와 거의 같은 방향이다. 그렇다면 저자성의 핵심을 ‘**creative control**’로 보면 어떨까 2025년 *Computer Law & Security Review* 논문은 인간-AI 공동창작에서 저자성을 창작 전 과정에 대한 인간의 창조적 통제와 연결하는 이론을 제안한다. 즉: **WHO CONTROLS THE CREATIVE TRAJECTORY?**가 중요하다. **HARD LOCK CREATOR ≠ AUTHOR ≠ OWNER ≠ RESPONSIBLE PARTY**. 네 개는 다르다. 아이디어의 기원과 소유도 다르다 아이디어 origin: AI 출력. 하지만:

- 내가 이해
- 발전
- 검증
- 다른 생각과 통합

하면 그 아이디어는: **내 사고체계의 일부**가 될 수 있다. 여기서는 세 단계를 분리해야 한다 **ORIGIN** 처음 어디서 나왔는가? **INTEGRATION** 누가 자신의 사고체계에 통합했는가? **DEVELOPMENT**

누가 그것을 확장·검증·결과로 만들었는가? 같은 사람이 아닐 수 있다. 하지만 출처를 지우면 안 된다 분산생산을 인정한다고: **provenance** 가 사라지는 것은 아니다. 오히려 AI 시대에는: 무엇이 어디서 왔는가 를 추적하는 것이 더 중요하다. 하지만 AI가 기억을 요약·재구성하면? 더 어려워진다. 내 과거: A. AI 요약: B. 나는 B를 읽고: 내가 원래 이렇게 생각했지. 라고 기억. 그러면: **memory retrieval** 이 아니라 **memory reconstruction**. 이 된다. 따라서 기억 AI에는 새로운 원칙이 필요하다 **MEMORY SUPPORT ≠ IDENTITY PRESCRIPTION**. AI는: 과거 기록에는 이렇게 남아 있습니다. 와 당신은 이런 사람입니다. 를 구분해야 한다. 이것이 **responsibility gap** 문제다 분산시스템에서는: **인과적 기여는 많고 최종 책임자는 불명확** 해질 수 있다. 그러나 2026년 저자성 논문은 이런 gap이 AI에게 저자성을 줘야 한다는 뜻이 아니라, 오히려 **인간과 기관의 책임구조를 더 정밀하게 설계해야 한다는 뜻**이라고 주장한다. **저자성도 비슷하다 AUTHORSHIP FOLLOWS MEANINGFUL CONTROL + ENDORSEMENT + ANSWERABILITY**. 이 세 축이 현재 가장 강한 후보다. **AI가 90% 문장을 썼지만 인간이 6축에서 강하다** 저자성: 상당히 유지 가능. 즉 한 단어 '주체'가 너무 많은 것을 숨긴다 최소: **CAUSAL AGENT COGNITIVE AGENT AUTHOR MORAL AGENT LEGAL PERSON CONSCIOUS SUBJECT**

를 분리해야 한다. 따라서 AI가 처음 제안한 생각도 내가:

- 이해
- 검증
- 반례검토
- 재구성

하면 내 사고의 일부 가 될 수 있다. 그러나: 출력 그대로 복사하고 의미도 모르면 **borrowed output** 에 가깝다. 반대로 AI와 토론하면서 학생이:

- 반론 제기
- 논지 수정
- 자기 예시 추가
- 최종 논증 구성

했다면 AI 사용에도 불구하고: **intellectual authorship** 은 강할 수 있다. **AI CONTRIBUTION DEPTH D1 — SURFACE** 문법·형식. **D2 — SYNTHESIS** 요약·구조. **D3 — REASONING** 논증·해석. **D4 — DISCOVERY** 가설·핵심아이디어. **D5 — DECISION** 무엇을 주장할지 결정. 깊어질수록: **human authorship requires stronger active oversight**. 이제 저자성과 정체성을 연결해보자 내가 10년간 AI와 함께 글을 쓴다. AI가: 내 문체. 주제. 생각. 과거 자료. 전부 기억. 점점 나보다: **나답게 글을 쓴다**. 그러면? 왜? 앞 장: 개인정체성에는

- 생물학적 연속성
- 심리적 연속성
- 인과적 역사

등 복합문제가 있다. 문체는 그중: **매우 작은 일부**. 따라서 질문이 바뀐다 AI가 내 생각을 훔쳤나? 보다 **AI와의 상호작용이 어떤 종류의 사람을 만들어가고 있는가?** 가 더 중요할 수 있다. 그렇다면 AI와 함께

형성된 나도 '진짜 나'인가 왜 아니겠는가? 인간은 원래:

- 부모
- 언어
- 책
- 학교
- 친구

로 형성된다. AI만 유독: 외부영향이라서 가짜 라고 할 이유는 없다. 아주 강한 문장 좋은 개인화 AI는 나를 잘 기억해야 하지만, 나를 과거의 나에 가두어서는 안 된다. 앞 장·29의 결과를 정체성 수준에서 압축한다. 이때 AI는 무엇인가 상황에 따라:

- 도구
- 공동생성기
- 비평가
- 검색자
- 기억보조자
- 아이디어 촉발자

일 수 있다. 하지만 현재 구조에서는: ACCOUNTABLE AUTHOR 와는 구별된다. HARD LOCKS v3.0 기여율 ≠ 저자성 생성 ≠ 저자성 창작성 ≠ 책임 저자 ≠ 소유자 저자 ≠ 최초 아이디어 발생자 AI 사용 공개 ≠ 저자성 포기 최종 승인 ≠ 자동적 실질통제 분산된 인과 ≠ 책임 소멸 human approval ≠ meaningful human control 문체 복제 ≠ 개인정체성 복제 기능적 결합행위자 ≠ 도덕적 인격 기능적 결합 ≠ 하나의 의식 AI 영향 ≠ 비진정한 자아 개인화 기억 ≠ 정체성 결정권 이 장의 가장 강한 결론 AI 시대의 저자성은 '누가 문장을 만들었는가'보다 '누가 의미를 선택하고, 결과를 이해하며, 자기 이름으로 승인하고, 틀렸을 때 고칠 책임을 지는가'의 문제로 이동할 가능성이 크다. 즉: AUTHORSHIP CONTROL + INTEGRATION + ENDORSEMENT + ANSWERABILITY. 기여도는 중요하지만 충분하지 않다. 그러면 AI와 함께 만든 생각도 충분히 '내 것'일 수 있다 단: 그것을 실제로 내가 소화했다면. 반대로: 내가 직접 타이핑했어도 이해하지 못한 생각 이라면 '내 것'이라는 주장도 약해진다. 반대로 AI 수를 보고 기사: 왜 좋은지 이해. 새 패턴 학습. 나중 AI 없이 재현. 그러면: AI-originated insight → human-integrated knowledge. 이 된다. 그래서 AI 시대의 가장 중요한 교육 질문도 바뀐다 누가 답을 만들었는가? 만으로는 부족하다. 더 좋은 질문: 누가 그 답을 이해하고 책임질 수 있는가? 이다. 그리고 이제 한 단계 더 깊은 문제가 열린다 인간과 AI가:

- 기억 공유
- 장기 계획 공유
- 가치 모델 공유
- 의사결정 공유

를 계속하면, 어느 순간: 하나의 더 큰 인지시스템 이라고 보는 것이 더 정확할 수도 있다. 그렇다면: 개인은 어디서 끝나는가? 앞 장의 self-model, 앞 장의 identity, 앞 장의 extended cognition이 모두 만난다. 나는 어디까지가 '나'인가 뇌 · 몸 · 기억 · 스마트폰 · AI · 확장된 마음 · 경계 · 개인정체성

피부 바깥까지 인지가 확장된다면 자아도 확장되는가 가장 중요한 시작점은 이것이다. **COGNITIVE BOUNDARY ≠ BIOLOGICAL BOUNDARY ≠ PHENOMENOLOGICAL BOUNDARY ≠ RESPONSIBILITY BOUNDARY ≠ LEGAL PERSON BOUNDARY**. 이 구분을 하지 않으면 스마트폰도 내 마음의 일부다에서 곧바로 스마트폰도 나다로 점프하게 된다. 확장된 마음 계열의 핵심 주장은 환경의 일부가 단순한 도구를 넘어 인지과정 자체의 **구성요소**가 될 수 있다는 것이다. 반면 embodied cognition 문헌은 몸의 역할, embedded cognition은 환경 의존성, extended cognition은 아예 인지시스템의 경계가 유기체 밖까지 뻗을 수 있다는 더 강한 주장을 구분한다. **가장 쉬운 사례부터 시작하자** 사람 A. 친구 전화번호를 기억한다. 머릿속에서 꺼낸다. 사람 B. 전화번호를 기억하지 않는다. 하지만 스마트폰 연락처를 항상 사용한다. 둘 다: **필요한 순간 전화번호를 이용한다**. 그렇다면: A의 기억은 내부 기억. B의 연락처는: **단순 도구인가, 아니면 기억체계의 외부 일부인가?** 이 질문이 extended mind의 출발점이다. **고전적 직관** 만약 외부 자원이:

- 항상 접근 가능하고
- 안정적으로 신뢰되며
- 별도 복잡한 판단 없이 사용되고
- 실제 행동을 지속적으로 안내한다면

그 자원을: **cognitive system의 일부**로 볼 수 있지 않겠는가? 즉: **WHERE COGNITION HAPPENS**를 **SKIN AND SKULL**로 미리 고정하지 말자는 주장이다. **중요한 HARD LOCK COGNITIVE DEPENDENCE ≠ COGNITIVE CONSTITUTION**. 이다. 안경 없으면 잘 못 본다. 그렇다고: **안경이 시각피질**인 것은 아니다. 단순 의존성과 시스템 구성성은 구분해야 한다. **하지만 스마트폰: 나?** 여기서 멈춘다. **NO**. 적어도 바라는 아니다. 인지과정 일부를 구성할 수 있다는 주장과: **PERSON IDENTITY**는 다른 문제다. **따라서 첫 번째 큰 분기 MIND EXTENSION과 PERSON EXTENSION**은 다르다. 내 사고과정이 외부에 걸쳐 있다고 해서: **내 생물학적 개체가 외부까지 늘어난다**는 뜻은 아니다. **생물학적 경계** 가장 직관적인 나: **피부 안**. 하지만 이것도 완전히 단순하지 않다. 미생물총. 의료기기. 인공관절. 이식장기. 경계 사례가 많다. 그래도 기본적으로 유기체의 지속성에는: **하나의 생명과정**이라는 강한 기준이 있다. Animalism은 바로 이 점을 강조한다. **이제 AI를 넣으면 확장인지가 훨씬 어려워진다** 노트북 메모: 내가 쓴 내용. 그대로 반환. AI:

- 내 기록 읽음
- 요약
- 추론
- 새 아이디어 생성

즉: **PASSIVE EXTERNAL MEMORY**가 아니라 **ACTIVE TRANSFORMATIVE PARTNER**. 가 된다. **이것이 2025년 LLM extended-mind 연구의 핵심 문제다** Smart·Clowes·Clark는 RAG 기반 'Digital Andy'를 분석하면서 LLM이 외부 기억과 결합될 경우 extended cognition의 흥미로운 후보가 될 수 있다고 검토했다. 하지만 그 사례에서는 LLM 자체가 외부기억 검색을 능동적으로 시작하지 않았다는 이유 등으로 **완전한 extended AI system**이라고 보기는 어렵다는 부정적 결론을 내렸다. 즉: **EXTERNAL MEMORY CONNECTED ≠ EXTENDED COGNITIVE SYSTEM AUTOMATICALLY**. 따라서 **PHENOMENOLOGICAL SELF** 지금 이 경험이 누구에게 나타나는가? 라는 경계다. 이진: **cognition** 보다 좁을 수 있다. 하지만 **agency**가 분산돼도 책임은 자동으로

같이 분산되지 않는다 앞 장: DISTRIBUTED CAUSATION ≠ NO ACCOUNTABILITY. 그리고: FUNCTIONAL AGENT ≠ MORAL PERSON. 유지. 예 AI가 90% 판단. 의사가 최종 승인. 그런데 의사에게:

- AI 근거 접근권
- 거부권
- 검토시간

이 없다. 그러면 단순: 인간이 버튼 눌렀으니 책임. 은 부당할 수 있다. **LEGAL PERSON BOUNDARY** 법적으로는 더 분명하다. 스마트폰이 내 인지 일부처럼 작동한다고 해서: **내 스마트폰이 법적 나**가 되는 것은 아니다. 내 AI가 계약을 추천해도: AI가 자동으로: **법적 당사자**가 되는 것도 아니다. **이 모델로 스마트폰을 검사하자** *Biological X. Cognitive* 일부 기능에서 O 후보. *Phenomenological X. Responsibility* 직접적 책임주체 아님. *Legal person X.* 즉: **COGNITIVELY CLOSE BUT PERSONALLY DISTINCT.** 인간의 몸은 어떨까 몸: *Biological O. Cognitive* 매우 강함. *Phenomenological* 몸감각을 통해 강하게 포함. *Responsibility* 행동의 실행기반. *Legal* 개인의 신체. 즉 몸에서는 다섯 경계가 훨씬 많이 겹친다. 이것이 embodied cognition이 중요한 이유다. 몸은 단순 입력장치가 아니라 인지의 구성에 깊게 기여할 수 있다. **AI가 이 겹침을 처음으로 풀어놓을 수 있다** 기억: 밖. 계획:밖. 추론: 밖. 그런데: 경험: 아직 안. 법적 인격: 인간. 즉: **FUNCTIONAL SELF BOUNDARIES BEGIN TO DIVERGE.** 이것이 AI 시대의 독특한 상황이다. 그래서 **self-model도 확장될 수 있다** 사람이 말한다. 나는 영어 못해. 번역 AI 있음. 실제 행동능력: 외국인과 협업 가능. 그러면: **internal skill은 낮지만 effective capability는 높다.** ‘나는 영어를 할 수 있는가?’ 자체가 애매해진다. 하지만 이게 곧 도구: 사람은 아니다 다시: **CAPACITY EXTENSION ≠ IDENTITY EXTENSION.** 이 구분이 핵심이다. 여기서 엄청나게 중요한 문제가 생긴다 **EXTENDED MIND**

WITHOUT EXTENDED CONTROL. 내 마음의 기능 일부가 외부에 있는데, 그 외부를 내가 통제하지 못한다. 그러면: **cognitive sovereignty risk.** 따라서 확장인지에는 권리문제가 붙는다 만약 외부자원이: 실제 인지기능의 핵심 일부 라면 그 자원에 대한:

- 접근
- 이동성
- 삭제
- 백업
- 수정

통제는 단순 소비자 편의보다 중요해질 수 있다. 이것은 재산권과도 완전히 같지 않다 내 안경을 소유. 내 스마트폰 소유. 하지만 클라우드 AI는: 구독. 데이터는 서버. 모델도 타인 소유. 그러면: **cognitive dependence** 와 **property/control rights** 가 갈라진다. 하지만 **SOCIAL SELF ≠ NUMERICAL SELF** 사람들이 나를 잘못 기억해도: 내가 다른 인간이 되는 것은 아니다. 다시: **identity layer separation.** 그리고 다섯 번째 **NORMATIVE SELF** 누가: 약속 책임 권리 를 갖는가. 이는 법적·도덕적 층이다. 매우 중요한 **HARD LOCK MULTIPLE SELF-DIMENSIONS ≠ MULTIPLE PEOPLE INSIDE ONE PERSON.** 이다. 다층적이라고 다중인격을 뜻하는 것이 아니다. 분석축이 여러 개라는 뜻이다. 예를 들어 조직에서도 병원 원장이: 병원 의사결정시스템의 핵심 일부. 그렇다고:

원장: 병원 은 아니다. 같이: AI+인간 시스템이 하나의 기능적 의사결정단위라도 **인간: 전체 시스템** 은 아니다. 그러면 **extended cognitive component**는 교체돼도 되는가 YES 가능. 수첩 A → 수첩 B. 스마트폰 A → B. 정보·기능 연속성 유지. 즉: **substrate replacement** 에도 cognitive extension은 지속될 수 있다. 따라서 **self-boundary**에는 '결합강도'도 필요하다 **WEAK COUPLING** 가끔 사용하는 검색엔진. **MEDIUM COUPLING** 매일 쓰는 스마트폰 기억. **STRONG COUPLING** 항상 켜져 있고 내 기억·계획·행동을 지속적으로 조절하는 AI. 이 강도가 증가하면: **cognitive extension claim** 은 강해진다. 따라서 프로젝트의 **SELF BOUNDARY MODEL v1.0**

S1 — ORGANISM 생물학적 자기유지 경계. *S2 — EMBODIED COGNITION* 몸 전체가 인지에 기여. *S3 — EXTENDED COGNITION* 외부도구까지 기능적 인지경계 확장. *S4 — PHENOMENAL SUBJECT* 경험이 나타나는 주체. *S5 — NARRATIVE / SOCIAL IDENTITY* 기억·역할·자기서사. *S6 — NORMATIVE / LEGAL PERSON* 책임·권리의 주체. 한 사람이 이 여섯 층을 동시에 가질 수 있다. 따라서 'AI가 내 일부가 된다'는 문장은 너무 모호하다 반드시 묻는다. **IN WHAT SENSE?** 인지? 기억? 정체성? 경험? 책임? 법? 따라서 **privacy**도 **self-boundary**와 연결된다 사진. 대화. 과거 판단. AI 기억. 이게 내 미래 의사결정에 계속 사용된다면: **informational privacy**가 **cognitive autonomy**와 연결된다. 그래서 **외부기억**은 단순 저장소가 아니다 이 될 수 있다. 단: **narrative self** 차원에서다. 생물학적 동일성까지 바뀌는 것은 아니다. 그럼 **인간의 독립성**은 끝나는가 아니다. 앞 장: **autonomy ≠ independence**. 오히려 앞으로 필요한 것은: **BOUNDED DEPENDENCE WITH RETAINED CONTROL**. 즉: 의존하지만

- 이동 가능
- 검증 가능
- 거부 가능
- 복구 가능.

따라서 **extended mind**는 정치철학으로 넘어간다 처음에는: **마음은 어디까지인가?** 였는데 이제: **확장된 마음의 인프라**를 누가 소유하는가? 가 된다. 이것은 엄청나게 현실적이다. 이것은 앞 장의 **preference shaping**보다 더 깊다 추천: 내 선택에 영향. 기억·계획 인프라: 무엇을 생각할 수 있는지 자체에 영향. 즉: **CHOICE ARCHITECTURE**에서 **COGNITIVE ARCHITECTURE**로 내려간다. **COGNITIVE SOVEREIGNTY v0.2** *C1 — ACCESS* 내 외부 인지자원에 접근할 수 있다. *C2 — PORTABILITY* 한 시스템에서 다른 시스템으로 옮길 수 있다. *C3 — INSPECTABILITY* 무엇이 저장·추론되는지 볼 수 있다. *C4 — CORRECTABILITY* 잘못된 기억·모델을 수정할 수 있다. *C5 — REVOCABILITY* AI 권한을 끊을 수 있다. *C6 — FALLBACK* AI 없이 최소기능을 복구할 수 있다. *C7 — PLURALITY* 한 제공자에게 완전히 잠기지 않을 대안이 있다. 이걸 AI 시대의 강한 프로젝트 가설로 남긴다. **복제가 특히 문제인 이유** 내 기억을 AI A에 복사. 동시에 AI B에도 복사. 둘 다: 내가 원본이다. 라고 함. 둘이 모두: **수적으로 동일한 나** 일 수 있는가? 일반적인 동일성 논리는 어려움을 겪는다. 이것은 **AI 디지털 불멸 논의도 막아준다** 내 기록으로 챗봇 생성. 그 챗봇이: **나를 잘 흉내** 낼 수 있다. 하지만: **내가 계속 살아 있다**는 주장은 훨씬 강하다. 이 장의 가장 큰 발견이 여기에 있다 우리는 계속: “나”를 하나의 것처럼 사용했다. 그러나 실제로는: **여러 종류의 연속성과 경계**가 한 사람에게 겹쳐 있다. AI는 이 겹침을 분리해서 보이게 한다. 이것은 철학적 혼란이 아니라 분석기회다 AI가: **기존에 숨어 있던 구분**을 드러낸다. 마치 바둑 AI가: **인간 직관 속에 숨은 평가함수**

를 드러냈듯, 인간-AI 결합은: 자아 안에 숨은 여러 경계 를 드러낸다. 반대로 가능하다 내부 기억하던 것을: AI에 저장. **externalization**. 따라서 인간 인지는: **INTERNALIZE ↔ EXTERNALIZE** 를 반복한다. 그렇다면 마음의 진짜 경계는 계속 움직이는가 기능적 인지 수준에서는: **YES, potentially**. 상황마다: 계산기 사용. AI 사용. 친구와 토론. 인지 네트워크가 바뀐다. 즉: **COGNITIVE BOUNDARY CAN BE TASK-DEPENDENT**. 따라서 질문 AI는 나의 일부인가? 의 답은: 어떤 의미에서는 **YES**, 다른 의미에서는 **NO**, 일부에서는 **UNKNOWN**. 이다. 이게 회피가 아니라 가장 정확한 답이다. **HARD LOCKS v3.1** 인지적 의존 ≠ 인지적 구성 인지적 확장 ≠ 생물학적 확장 확장된 마음 ≠ 확장된 의식 인지적 파트너 ≠ 동일한 사람 기능적 agency ≠ 법적 인격 기억 외부화 ≠ 개인정체성 외부화 전체 디지털 복제 ≠ 개인 생존 심리적 유사성 ≠ 수적 동일성 AI가 나를 기억함 ≠ AI가 나임 AI와 결합 ≠ 하나의 의식 자율성 ≠ 기술독립 확장인지 ≠ 외부인프라에 대한 통제권 불필요 이 장의 가장 강한 결론 '나의 경계는 하나가 아닐 수 있다. 생물학적으로 나는: 하나의 유기체. 인지적으로 나는: 몸과 외부도구를 포함하는 더 넓은 시스템에 참여할 수 있다. 현상적으로 나는: 경험의 중심. 사회적으로 나는: 관계와 기억 속의 인물. 규범적으로 나는: 책임과 권리의 주체. AI는 이 경계들 가운데 특히: **cognitive narrative agency** 경계를 크게 흔들 수 있지만, 모든 경계를 동시에 지우지는 않는다. 따라서 미래의 자유 문제도 바뀐다 고전적 자유: 국가가 내 행동을 제한하는가? AI 시대에는 추가:

내 기억·정보·판단을 가능하게 하는 인지 인프라를 누가 설계하고 통제하는가? 즉: **BEHAVIORAL FREEDOM** 에서 **COGNITIVE SOVEREIGNTY** 까지 확장된다. 이것이 앞 장의 헌법과 앞 장의 인지주권이 만나는 지점이다 정치헌법: 권력 집중을 제한. AI 인프라: 인지권력 집중을 제한. 공통 문제: 누군가가 다른 사람의 미래 선택공간을 지나치게 통제하지 못하도록 하는 것. 그런데 이제 아주 큰 질문이 생긴다 개인은 이미:

- 언어
- 사회
- 책
- 인터넷
- AI

에 깊게 의존한다. 그렇다면: 인간 지능은 애초에 개인의 것이었는가? 혹시 인간의 가장 큰 지능은 처음부터: 개인 뇌 보다 여러 사람과 외부 기억을 연결한 집단적 인지체계에 있었던 것은 아닐까?

15장. 집단지능은 집단의식이 아니다

문명이 정보를 모으고 판단하고 행동할 수 있다면 하나의 거대한 마음인가?

집단행동, 분산인지, 집단지능, 집단행위, 공동의도, 집단의식은 서로 다른 주장이다. 문명은 하나의 뇌보다 인지제도와 오류수정 루프의 생태에 가깝다. 문명은 하나의 거대한 마음인가 집단지능 · 분산인지 · 공동의도 · 집단행위 · 시장 · 과학 · 인터넷 · AI 사회가 정보를 처리한다고 해서 사회가 '생각한다'고 말할 수 있는가 이 장의 핵심은 여섯 개를 분리하는 것이다. **COLLECTIVE BEHAVIOR ≠ DISTRIBUTED COGNITION ≠ COLLECTIVE INTELLIGENCE ≠ COLLECTIVE AGENCY ≠ COLLECTIVE INTENTIONALITY ≠ COLLECTIVE CONSCIOUSNESS**. 집단의도 철학에서도 공동의도는 개인들의 의도 단순 합으로 환원되지 않는다고 보는 입장이 강하지만, 동시에 그것을 참여자들 위에 떠 있는 별도의 '집단 정신'이 소유한다고 볼 필요는 없다는 것이 핵심 긴장이다. **경기장의 파도타기부터 시작하자** 관중 5만 명. 한쪽에서 사람들이 일어난다. 옆 사람이 따라 일어난다. 파도가 경기장을 돈다. 거대한 패턴. 누가 전체 파도를 설계했는가? 아무도 아닐 수 있다. 그렇다면: **집단행동**. 그러나: **집단이 생각했다?** 까지는 아니다. **첫 번째 층 COLLECTIVE BEHAVIOR** 많은 개체의 상호작용에서: **거시적 패턴** 이 나타난다. 예:

- 군중 이동
- 새떼
- 물고기 떼
- 개미 집단
- 금융시장 매수열풍

이런 현상은: **중앙두뇌 없이도** 발생할 수 있다. 따라서 **첫 번째 HARD LOCK COORDINATED PATTERN ≠ COLLECTIVE MIND**. 이다. 질서가 생겼다는 것만으로 마음이 생긴 것은 아니다. 앞 장의: **SELF-ORGANIZATION ≠ SELF-AWARENESS** 와 같은 오류다. **두 번째 층 DISTRIBUTED COGNITION** 이제 여러 사람이 실제로: 서로 다른 정보와 기능을 나눠 가진다. 예: 수술팀. 외과의. 마취과. 간호사. 모니터. 프로토콜. 누구 하나가: 전체 상태를 완전히 머릿속에 가지고 있지 않을 수 있다. 하지만 팀 전체는: **수술을 수행**. 여기서는 **사고과정이 실제로 분산된다** 한 사람이: 혈압 감시. 다른 사람: 출혈 확인. 다른 사람: 기구 준비. 기록시스템: 약물시간 저장. 즉: **MEMORY PERCEPTION ACTION** 이 여러 요소에 분산된다. 이런 구조는 단순 '여러 사람이 함께 있음'보다 훨씬 강하다. 따라서 **distributed cognition의 핵심 TASK-RELEVANT COGNITIVE FUNCTION IS DISTRIBUTED ACROSS PEOPLE + ARTEFACTS + PROCEDURES**. 이것은 충분히 강한 설명범주다. 하지만 여전히 **DISTRIBUTED COGNITION ≠ ONE CONSCIOUS SUBJECT**. 수술팀 전체가: 하나의 통증을 느끼지 않는다. 각 구성원은 각자의 경험을 가진다. **세 번째 층 COLLECTIVE INTELLIGENCE** 집단이: **개인보다 문제를 더 잘 푸는가?** 예: 여러 독립 추정치를 평균. 오차 상쇄. 개인보다 정확. 이것이: **wisdom of crowds**. 하지만 **집단지능은 단순 평균이 아니다** 현대 집단지능 연구는 단순 다수결보다:

- 서로 다른 정보
- 독립적 판단
- 정보공유 방식
- 네트워크 구조
- 의견통합 규칙

이 성능에 결정적이라고 본다. *Nature Reviews Psychology* 리뷰도 집단지능을 단순 '사람이 많으면 더 똑똑하다'가 아니라 정보집계와 사회적 학습 메커니즘의 문제로 다룬다. **집단은 개인보다 더 똑똑할 수 있다** 가능. 그러나 반대도 가능. **HERDING. GROUPTHINK. INFORMATION CASCADE.** 다수가: **같은 오류** 를 복제할 수 있다. 따라서: **GROUP SIZE $\uparrow$ $\neq$ INTELLIGENCE $\uparrow$.** **집단지능의 핵심조건 후보** *CI1 — DIVERSITY* 서로 다른 정보·가설. *CI2 — PARTIAL INDEPENDENCE* 오류가 완전히 같은 방향으로 동조되지 않음. *CI3 — INFORMATION TRANSMISSION* 중요한 정보가 전달됨. *CI4 — AGGREGATION* 정보를 통합할 메커니즘. *CI5 — ERROR CORRECTION* 틀린 집단판단을 수정할 수 있음. *CI6 — MEMORY* 배운 것을 보존. 이 구조가 앞 장 문명학습과 정확히 겹친다. 즉 **문명은 집단지능을 가질 수 있다** 과학. 시장. 법. 행정. 모두: **개인이 혼자 처리할 수 없는 규모의 정보** 를 분산처리한다. 따라서: **CIVILIZATION HAS COLLECTIVE INTELLIGENCE** 라는 약한 의미의 문장은 충분히 방어 가능하다. 하지만 '지능' 자체도 여러 의미다 문제해결. 적응. 예측. 학습. 따라서: **COLLECTIVE INTELLIGENCE** 라고 할 때: 어떤 수행능력? 을 반드시 묻는다. 예 — **시장** 수백만 명:

- 수요
- 비용
- 기대
- 위험

정보를 행동에 반영. 가격 형성. 누구도: **모든 정보를 알지 못함.** 그런데 가격은: **분산정보 일부를 압축한 신호.** 그러면 시장은 생각하는가 약한 의미: **INFORMATION AGGREGATION SYSTEM.** YES. 강한 의미: **시장이 하나의 주체로 생각한다.** NO basis. 시장은 무엇을 '안다'고 말할 수 있는가 주의. "시장이 알고 있다." 는 표현은: **useful shorthand** 일 수 있다. 하지만: 시장 전체가: 나는 금리가 오를 거라고 믿는다. 라는 하나의 신념상태를 가진다고 바로 볼 수는 없다. **앞 장의 정보 구분을 적용하면** 가격: **SIGNAL.** 분산된 정보를 반영할 수 있다. 그러나: **PERFECT REPRESENTATION** 은 아니다. 시장:

- 투기
- 정보비대칭
- 독점
- 감정

으로 왜곡 가능. 따라서: **AGGREGATION $\neq$ TRUTH.** 네 번째 층 **COLLECTIVE AGENCY** 이제 더 강한 질문. 회사가: 공장을 짓기로 결정했다. 우리는 자연스럽게 말한다. "**회사가 결정했다.**" 회사에 정말 agency를 귀속할 수 있는가? 예 — **주식회사** 이사회 투표. CEO 실행. 법인 계약. 어느 한 개인이: **회사 전체 의사** 와 동일하지 않다. CEO가 반대해도 이사회 결정에 따라 회사가 행동할 수 있다. 이

경우 조직수준 설명이 유용하다 “직원 10,347명의 개별 움직임.” 보다: “회사가 인수했다.” 가 더 좋은 설명이다. 즉: COLLECTIVE AGENCY CAN BE EXPLANATORILY REAL. 하지만 이것도 집단의를 뜻하지 않는다 회사:

- 계약
- 계획
- 전략

가능. 하지만: 회사가 아침에 우울감을 느낀다 는 주장은 완전히 다른 종류다. 따라서 HARD LOCK GROUP AGENCY ≠ GROUP EXPERIENCE. 이다. 다섯 번째 층 COLLECTIVE INTENTIONALITY 두 사람이: 각자 산책. 우연히 같은 길. vs 두 사람이: 같이 산책하자. 하고 함께 걷는다. 외형: 비슷. 구조: 다르다. 공동의도에는 ‘우리’ 구조가 있다 단순: 내 목표 A + 네 목표 A 가 아니다. WE INTEND TO A. 공동행동 철학은 바로 이 차이를 분석한다. 같은 장소로 뛰어나는 군중과 함께 안무를 수행하는 무용수 집단은 물리적 행동이 비슷해도 shared agency 여부가 다르다. 중요한 점 집단의도는: 단순 합계 가 아닐 수 있다. 하지만 그 의도가 반드시: 별도 초개인적 마음 에 존재해야 하는 것도 아니다. SEP는 이 긴장을 다음처럼 정리한다. 집단의도는 개인의도 단순 합보다 더 강하다. 그러나 의도는 여전히 참여자들이 갖는다. 따라서 EMERGENT SOCIAL STRUCTURE 와 MYSTICAL GROUP MIND 를 구분할 수 있다. 우리가 함께:

- 약속하고
- 역할 나누고
- 서로 기대

하면: 새로운 집단적 행동구조 가 생긴다. 그렇다고: 제3의 의식주체 가 머리 위에 생성될 필요는 없다. 예 — 오케스트라 각 연주자: 개별 의식. 하지만: 전체 연주는: 어느 한 명의 결과가 아니다. 지휘자도: 모든 소리를 직접 생산하지 않는다. 그런데: 오케스트라가 연주한다. 라는 문장은 문자 그대로 유용하다. 그래서 collective agency를 인정한다고 개인주의를 완전히 버릴 필요는 없다 가능한 구조: INDIVIDUAL MINDS + RELATIONAL ORGANIZATION → COLLECTIVE AGENCY. 즉: GROUP-LEVEL PROPERTY WITHOUT GROUP-LEVEL CONSCIOUS SUBSTANCE. 이것이 현재 가장 안전하다. 그런데 문명 전체도 agent인가 여기서는 훨씬 어렵다. 회사:

- 경계
- 목표
- 의사결정 절차
- 대표자

가 있다. 문명 전체: 어디가 경계? 누가 결정? 무슨 목표? 불분명. 따라서 ORGANIZATION 와 CIVILIZATION 을 동일시하면 안 된다. 회사·국가: stronger collective agency candidates. 인류 전체: much weaker. 인류 전체가 기후문제에 대응한다고 하자 국가. 기업. 과학자. 시민. 서로 갈등. 단일 목표 없음. 그러면: HUMANITY WANTS X 라고 말하기 어렵다. 즉 집단행위에도 경계가 필요하다 PROJECT COLLECTIVE AGENCY TEST: A1 — MEMBERSHIP 누가 집단인가? A2 — DECISION RULE 무엇이 집단결정인가? A3 — GOAL STRUCTURE 공동목표가 있는가? A4 — REPRESENTATION 누가 집단을 대표하는가? A5 — IMPLEMENTATION 결정을 행동으로 전환할 수

있는가? A6 — PERSISTENCE 구성원교체 후에도 행위주체가 지속되는가? 회사·국가: 강함. ‘문명 전체’: 훨씬 약함. **여섯 번째 층 COLLECTIVE CONSCIOUSNESS** 여기서 가장 큰 점프가 나온다. 사람들이 연결. 정보 공유. 집단이 행동. 그러면: **하나의 집단적 경험** 이 생기는가? 현재: NO ESTABLISHED BASIS. 무엇이 부족한가 개인 의식에서도: **경험을 만드는 충분조건** 을 아직 모른다. 앞장에서 확인했다. 그런데 사회 전체에: **consciousness** 를 귀속하려면 더 강한 증거가 필요하다. **단순 연결로는 부족하다** 뉴런이 연결돼 있다. 의식과 관련. 사람도 인터넷으로 연결. 그러므로: **인터넷도 의식?** 이진: SAME FORM → SAME MECHANISM 오류다. 앞 장의 HARD LOCK 그대로다. 뉴런과 인간은 노드라는 점만 같다 하지만:

- 신호속도
- 결합강도
- 조직
- 가소성
- 생물학
- 기능

전부 다르다. 따라서: BRAIN NETWORK ≠ SOCIAL NETWORK. 매우 중요한 HARD LOCK NETWORK TOPOLOGY SIMILARITY ≠ CONSCIOUSNESS EQUIVALENCE. 이다. 따라서 사회가 더 큰 네트워크라는 사실은 거의 아무것도 증명하지 않는다 SIZE ≠ SUBJECTIVITY. CONNECTIVITY ≠ EXPERIENCE. 그런데 ‘collective attention’은 가능한가 여기는 흥미롭다. 사회가: 어떤 사건에: 뉴스 검색 자원 토론 을 집중. 우리는: 사회의 관심이 이쪽으로 쏠렸다. 라고 말한다. **이 표현은 완전히 비어 있지 않다** 2026년 Synthese 논문은 ‘집단적 주의’를 형이상학적으로 하나의 집단 마음이 존재한다는 주장보다, 사회가 자원·행동을 특정 사안에 선택적으로 배분하는 현상을 설명하는 실용적 개념으로 보자고 제안한다. 즉: COLLECTIVE ATTENTION WITHOUT COLLECTIVE CONSCIOUSNESS 가 가능하다. 예 — 전쟁 발생 사회적 attention: 급증. 뉴스. 정책. 검색. 돈. 사람. 특정 문제에 집중. 이것은: **resource allocation phenomenon**. 굳이: **문명이 그 전쟁을 ‘느낀다’** 고 말할 필요 없다. 이 구분이 아주 좋다 개인의 attention: 경험과 연결될 수 있음. 집단 attention: **collective resource allocation**. 같은 단어지만: **implementation differs**. 과학은 집단지능의 가장 강한 사례 중 하나다 한 과학자는: 전체 과학을 모른다. 하지만:

- 논문
- 데이터
- 실험
- 동료평가
- 학회
- 기록

을 통해: 지식이 집단적으로 축적. 과학공동체는 개인보다 오래 산다 과학자 죽음. 논문 남음. 다음 세대 수정. 즉: MEMORY OUTLIVES INDIVIDUALS. 앞 장의 institutional memory. **과학은 또 오류를 교정한다** 가설. 실험. 비판. 재현. 수정. 즉: COLLECTIVE LEARNING. 앞 장의 정의와 정확히 맞는다. 따라서 문명은 ‘학습한다’고 말할 수 있는가 조건부 YES. 우리가 학습을: 과거 상호작용이 보존되어 미래 판단·행동을 체계적으로 변화시키는 것 으로 정의했기 때문이다. 이

정의에서는: **institutions can learn**. 하지만 누가 배우는가 개인? 조직? 제도? 모두 가능. 예: 항공사 사고. 사고조사. 규정 변경. 새 조종사 교육. 사고 당시 사람은 퇴직. 그래도: **organization learned**. 이 표현은 매우 유용하다. 여기서 **learning**과 **consciousness**가 다시 분리된다 **LEARNING** ≠ **EXPERIENCE**. 회사도: 배울 수 있음. 회사 전체가: 사고를 슬퍼하는 하나의 의식을 가져야 할 필요는 없다. 시장도 학습하는가 가격·행동이 과거 결과 때문에 변화. 약한 기능적 의미: **possible**. 하지만: 시장 전체가 모델을 명시적으로 보유 한다고 할 때는 더 조심. 민주주의는? 선거결과. 정책실패. 정권교체. 법 개정. 이것도: **institutional learning candidate**. 하지만: 모든 정책변화: 학습은 아니다. 권력변화일 수도 있다. 앞 장 HARD LOCK 유지. AI가 들어오면 집단지능이 어떻게 변하는가 과거: human ↔ human. 이제: human ↔ AI ↔ human ↔ database ↔ AI agents. 즉: **HYBRID COLLECTIVE INTELLIGENCE**.

2026년 *Nature Machine Intelligence*는 집단지능을 고정 네트워크에서 이동하는 개체들의 집단까지 비교하며, 시스템마다 학습·정보전달·환경수정 방식이 크게 다르다고 강조한다. 즉 '집단지능'이라는 공통이름 아래에도 메커니즘은 매우 다양하다. 이것은 우리 프로젝트의 원칙과 완전히 맞는다. **SAME LABEL ≠ SAME MECHANISM**. LLM agent 집단도 등장한다 여러 AI agent가:

- 대화
- 협력
- 경쟁

하면서 집단행동을 만든다. 2026년 *npj Artificial Intelligence* 연구도 LLM 기반 인지에이전트 네트워크에서 집단적 행동이 어떻게 출현하는지 분석하기 시작했다. 그런데 **multi-agent AI**가 집단의식을 가진다는 뜻인가 NO. 현재 확인되는 것: **EMERGENT COLLECTIVE BEHAVIOR**. 가능. **COLLECTIVE CONSCIOUSNESS**. 전혀 다른 주장. 오히려 AI는 집단행동과 마음을 구분하는 좋은 실험대다 에이전트 100개. 상호작용. 사회적 규범 같은 패턴. 전체적으로 합리적 행동. 그렇다고: 101번째 마음 이 위에서 태어난다고 볼 이유는 없다. 이것이 사회에도 적용된다 개인 1억 명. 상호작용. 시장. 법. 언어. 문화. 거대한 질서. 그렇다고: 새로운 하나의 주관적 경험 이 필요하지 않다. 이제 언어를 보자 개인은: 언어 전체를 발명하지 않았다. 태어났을 때 이미 존재. 언어를 배우면: 수천 년간 축적된 구분과 개념 을 사용할 수 있다. 따라서 언어는 집단인지 인프라다 언어가 없으면: 각 세대가 모든 개념을 다시 만들어야 한다. 언어는: **COGNITIVE COMPRESSION ACROSS GENERATIONS**. 이다. 단어 하나에는 집단역사가 압축된다 '유전자'. '인플레이션'. '권리'. 한 단어를 배우면: 엄청난 이전 사고 를 호출할 수 있다. 즉: 라는 약한 표현은 강하다. 책은 시간축을 추가한다 죽은 사람의 생각. 다음 세대가 읽음. 비판. 확장. 따라서: **COGNITION CROSSES GENERATIONS**. 문명의 지능이 개인보다 강한가 가장 큰 이유 중 하나다. 인터넷은 공간·시간 비용을 낮춘다 정보접근. 협업. 검색. 거대한 기억망. 하지만: **information volume** ↑ 가 **collective intelligence** ↑ 를 보장하지 않는다. 앞 장 유지. 왜?

- 허위정보
- 중복
- 관심경쟁
- 극화
- echo chamber

때문. 즉: **CONNECTIVITY CAN AMPLIFY ERROR TOO**. 따라서 집단지능에도 오류구조가 필요하다 좋은 집단인지: 정보전파 빠름 만이 아니다. 필요: 오류 격리 반론 독립검증 다양성 보존. 바이러스처럼 오류도 퍼진다 사람 A 거짓정보. ↓ 100명 공유. ↓ 1만 명 믿음. 확산성 높음. 하지만: **truth value low**. 따라서: **INFORMATION TRANSMISSION EFFICIENCY ≠ EPISTEMIC QUALITY**. AI가 이 문제를 악화시킬 수도 있다 AI 생성 콘텐츠. 엄청난 양. 상호 인용. AI가 다시 학습. 잘못된 정보가: **synthetic consensus** 를 만들 수 있다. 반대로 AI는 집단지능을 높일 수도 있다

- 다양한 문헌 연결
- 반례 검색
- 번역
- 정보 접근

즉: **epistemic friction** ↓. 그 결과는: 설계와 제도에 달려 있다. 따라서 **HUMAN-AI CIVILIZATION**에서는 새 **bottleneck** 과거: **INFORMATION CREATION**. 미래: **VALIDATION PROVENANCE DIVERSITY TRUST CALIBRATION**. 앞 장의 결론이 집단수준으로 확장된다. 이제 **집단책임 문제** 회사가 환경오염. 누가 책임? 직원 한 명? CEO? 이사회? 회사? 집단책임 문헌에서는 개인주의적 입장과 집단 자체를 책임주체로 인정하는 입장이 모두 존재하며, 집단행위·집단의도와 연결된 별도 논쟁영역을 형성한다. 중요한 것은 **COLLECTIVE RESPONSIBILITY** 를 인정한다고: 개인 책임이 사라지는 것은 아니다. 둘 다 가능. 예 회사 차원의: **조직문화 인센티브 절차** 가 사고를 만들었다. 동시에 특정 관리자도: 명백히 잘못. 따라서: **SYSTEM RESPONSIBILITY + INDIVIDUAL RESPONSIBILITY**. 앞 장 그대로. 이것이 집단마음 비유가 위험한 이유 “사회가 원했다.” 라고 말하면: 누가 실제 권력을 행사했는지 숨길 수 있다. 예: 시장이 원했다. 국민이 원했다. 회사가 결정했다. 실제로는 특정 집단이 더 큰 권한. 따라서 집단수준 설명에도 **POWER MAP**이 필요하다 누가:

- agenda 설정
- 정보 선택
- 의사결정
- 실행

했는가? **COLLECTIVE OUTCOME ≠ EQUAL CONTRIBUTION**. 앞 장의 **social choice**가 돌아온다 “국민의 의지.” 하나의 마음처럼 표현. 그러나 실제: **diverse preferences + aggregation rule**. 이다. 따라서: **GENERAL WILL ≠ LITERAL SINGLE MIND**. 집단지능과 민주주의도 같은 것이 아니다 전문가 집단이: 사실예측에서 더 정확. 그러나: **political authority** 자동 획득? NO. 앞 장: **EPISTEMIC SUPERIORITY ≠ NORMATIVE SOVEREIGNTY**. 즉 집단수준에서도 네 종류의 **Correctness** 유지 과학: **truth**. 시장: **coordination**. 민주주의: **legitimacy**. 법: **rights/judgment**. 하나의: **SUPER-BRAIN** 으로 통합하면 기능 차이를 잃는다. 더 좋은 모델 과학. 시장. 국가. 법. 대학. 기업. 개인. AI. 각각 다른:

- 정보
- 목표
- 오류교정

을 갖는다. '생태계'도 비유이므로 조심한다 핵심 뜻: **multiple semi-autonomous systems interact**. 하나의 중앙통제자 없음. 완전 통합도 없음. 따라서 문명의 지능은 ONE INTELLIGENCE 보다 PLURAL INTELLIGENCE ARCHITECTURE. 가 더 정확하다. 앞 장의 MULTI-LOOP GOVERNANCE와 완전히 만난다 시장: 한 오류루프. 과학: 다른 오류루프. 법: 다른 오류루프. 민주주의: 다른 오류루프. 즉: CIVILIZATION LEARNS THROUGH MULTIPLE PARTIALLY INDEPENDENT LOOPS. 왜 여러 루프가 좋은가 하나가 틀리면: 다른 루프가 공격. 시장실패: 법. 정부오류: 선거·법원. 과학오류: 재현·비판. 즉: COGNITIVE CHECKS AND BALANCES. 앞 장의 권력분립이 인지론으로 돌아온다 CONCENTRATED COGNITION 도 위험할 수 있다. 한 기관이:

- 데이터
- 모델
- 판단
- 평가

전부 독점. 그러면: **epistemic single point of failure**. AI가 이것을 더 중요하게 만든다 초거대 AI 하나가:

- 검색
- 판단
- 추천

모두 장악. 효율 높음. 하지만: **diversity ↓ alternative models ↓ correction pathways ↓**. 즉: COGNITIVE CENTRALIZATION ↔ EPISTEMIC RESILIENCE. trade-off. 아주 중요한 프로젝트 가설 문명의 지능은 하나의 가장 똑똑한 중앙두뇌에서 나오기보다, 서로 다른 오류를 보는 여러 인지체계가 경쟁·협력·교정하는 구조에서 나올 수 있다. 이견 강하다. 이 구조를 바둑으로 비유하면 한 기사 혼자. vs 여러 기사 검토. AI 여러 모델. 다른 평가방식. 반론. 더 robust. 하지만: **group analysis always better** 는 아니다. 조율비용이 있다. 따라서 집단지능에도 비용이 있다

- 의사결정 느림
- 책임 분산
- 갈등
- 정보과부하

즉: DIVERSITY / REDUNDANCY ↔ SPEED / COORDINATION. 앞 장과 반복. 집단지능 최적화도 단일 최대가 아니다 너무 분산: 결정불능. 너무 중앙: 오류집중. 따라서: COORDINATION WITHOUT COGNITIVE MONOPOLY 가 강한 후보다. 이제 '문명은 하나의 마음인가?'에 답할 수 있다 약한 의미: 정보를 저장하고 처리하고 학습하는가? YES. 중간: 조직된 집단이 실제 행위주체가 될 수 있는가? YES, SOME CAN. 강한 의미: 인류 전체가 하나의 공동의도와 목적을 갖는가? GENERALLY NO. 가장 강한 의미: 문명이 하나의 주관적 경험을 하는가? NO EVIDENCE. 따라서 질문 자체를 바꾸는 편이 낫다 문명은 마음인가? 보다: 문명은 어떤 종류의 인지기능을 어느 구조에서 수행하는가? 이게 훨씬 생산적이다. PROJECT COLLECTIVE MIND LADDER v1.0 L0 — AGGREGATE 개별 행동 단순합. L1 — COORDINATED BEHAVIOR 상호작용 패턴. L2 — DISTRIBUTED

COGNITION 인지기능 분산. L3 — COLLECTIVE INTELLIGENCE 집단수준 문제해결·학습. L4 — COLLECTIVE AGENCY 집단수준 목표·행동. L5 — COLLECTIVE INTENTIONALITY 우리·의도·공동목표. L6 — COLLECTIVE CONSCIOUSNESS 하나의 경험주체. L0→L5는 사례별로 실제 가능. L6: NOT ESTABLISHED. 그리고: LOWER LEVEL DOES NOT AUTOMATICALLY GENERATE NEXT LEVEL. HARD LOCKS v3.2 집단행동 ≠ 집단지능 분산인지 ≠ 집단의식 집단지능 ≠ 집단행위 집단행위 ≠ 집단의식 공동의도 ≠ 초개인적 정신 조직 학습 ≠ 조직 경험 시장 정보집계 ≠ 시장이 생각함

사회적 관심집중 ≠ 사회적 의식 네트워크 연결성 ≠ consciousness 큰 네트워크 ≠ 큰 마음 집단결과 ≠ 모두의 동일한 의지 collective responsibility ≠ individual responsibility disappearance 집단지능 ≠ 정치적 정당성 문명 ≠ 거대한 뇌 이 장의 가장 강한 결론 문명은 하나의 거대한 마음이라기보다, 개인·조직·제도·기술이 서로 다른 인지기능을 나누어 수행하고 서로의 오류를 교정하는 다층적 집단인지 체계로 보는 편이 더 강하다. 즉: COLLECTIVE COGNITION WITHOUT COLLECTIVE CONSCIOUSNESS. 이게 현재 정본이다. 그리고 AI는 이 구조를 엄청나게 바꿀 수 있다 지금까지 문명은: 인간들 사이의 집단지능. 앞으로: HUMAN + AI AGENTS + INSTITUTIONS + DATABASES 가 결합. 즉: HYBRID CIVILIZATIONAL COGNITION. 그런데 여기서 매우 큰 문제가 열린다 만약 인간 사회의 많은 판단을: AI agent들이 수행하고, 사람들이 그 결과를 따라가며, AI끼리 서로 정보를 교환한다면: 누가 문명의 목표를 결정하는가? 집단지능이 높아져도 인간 통제는 약해질 수 있다 예: AI 경제시스템. 생산성 최적. 물류 완벽. 정책 자동조정. 그런데 인간은: 왜 그런 결정이 내려졌는지 이해하지 못함. 그러면: CIVILIZATIONAL PERFORMANCE ↑ while HUMAN POLITICAL AGENCY ↓ 가능하다. 하지만 'AI를 줄이자'가 답은 아니다 도구를 거부하면: 집단지능 손실. 더 좋은 문제: 어떻게 AI를 문명적 인지체계에 넣되 인간이 목표설정·가치판단·정당성 구조의 주권을 잃지 않게 할 것인가? 이다. 즉 앞 장의 COGNITIVE SOVEREIGNTY가 집단수준으로 올라간다 개인:

cognitive sovereignty. 문명: COLLECTIVE COGNITIVE SOVEREIGNTY. 프로젝트 가설: 사회가 외부 AI 인지인프라에 깊게 의존하더라도 목표·규칙·검증·수정권을 정당한 인간 제도 아래 유지할 수 있는 능력. 이것은 기술정책 문제가 아니라 문명구조 문제가 된다 누가:

- 모델을 소유?
- 데이터 통제?
- 목표 설정?
- AI 행동 중단?
- 다른 모델 허용?

즉: AI GOVERNANCE COGNITIVE CONSTITUTIONAL DESIGN. 이라는 강한 연결이 나온다. 앞 장 헌법과 다시 합쳐진다 헌법이: 정치권력을 분산·제한 했다면, AI 시대에는: 인지권력 도 중요해진다. 누군가가: 무엇을 사람들이 보며 무엇을 믿으며 어떤 선택지를 보게 되는지 를 통제할 수 있기 때문이다. 그래서 다음 질문은 거의 필연적이다 문명은 하나의 마음이 아니다. 하지만: 매우 강한 집단인지 시스템 이 될 수 있다. 그리고 AI가 그 중심에 들어온다. 그러면: 문명이 더 똑똑해질수록 더 좋은 문명이 되는가? 정답은 분명하지 않다. 나치국가도 행정·기술·조직 능력이 높을 수 있었다.

지능과 선행은 다르다.

16장. 지능 이후의 지혜와 기억

더 똑똑한 문명이 더 좋은 문명이 아니라면 지혜로운 문명은 무엇을 할 수 있어야 하는가?

지능은 증폭기이며 방향은 별도다. 문명적 지혜는 현실접촉·가치복수성·자기제한·정당성·수정가능성을 조직하고, 과거를 현재의 행동규칙과 재평가 절차에 삽입하는 능력이다. 이 장에서는 더 똑똑함과 더 좋음 사이를 완전히 끊어놓는 게 핵심입니다. 2026년 새로 정리된 분석철학의 wisdom 문헌도 지혜를 단순 지식량이나 문제해결능력으로 보지 않고, 무엇이 중요한지 알고 여러 가치와 상황을 종합해 잘 판단하는 능력과 연결합니다. 더 똑똑한 문명은 더 좋은 문명인가 지능 · 지혜 · 가치 · 권력 · 효율 · 자유 · 문명적 목표 최선수를 잘 찾는 능력과 어떤 게임을 해야 하는지 아는 능력은 같은가 이 장의 핵심 구분은: INTELLIGENCE 무엇을 할 수 있는가? WISDOM 무엇이 중요하며, 어떤 상황에서 무엇을 해야 하는가? GOODNESS 그 선택이 실제로 정당하고 가치 있는가? 이다. 셋은 겹칠 수 있지만 동일하지 않다. 아주 유능한 독재국가를 생각하자 정부는:

- 경제예측 정확
- 세금징수 효율적
- 범죄추적 뛰어난
- 물류 완벽
- 감시시스템 정교

하다. 즉: 높은 행정지능. 그런데 동시에:

- 반대자 탄압
- 표현자유 없음
- 소수자 억압
- 정치권력 교체 불가능

하다. 질문: 더 똑똑한 국가인가? 어떤 기능에서는 YES. 더 좋은 국가인가? 별도 문제다. 이것만으로 첫 번째 결론이 나온다 INTELLIGENCE ≠ MORAL GOODNESS. 매우 영리한 범죄자. 뛰어난 전략가. 정교한 선전가. 모두 지능적일 수 있다. 그러나: 지능은 목표의 정당성을 보장하지 않는다. 바둑에서는 왜 이 문제가 잘 보이지 않았을까 바둑은 목표가 이미 주어진다. WIN. 그러므로 지능이 높아지면: 승률 ↑. 즉: intelligence 와 performance 가 거의 같은 방향으로 간다. 그래서 문명의 지능이 높아져도 무엇을 최적화할지 모르면: MORE POWERFUL OPTIMIZATION OF THE WRONG THING 이 가능하다. 앞 장부터 계속 반복된 문제다. 지능을 다시 정의해보자 이 프로젝트에서 약한 기능적 정의: INTELLIGENCE 환경에서 중요한 구조를 찾아 예측하고 선택하며 목표를 달성하는 능력. 이 정도는 충분히 강하다. Aristotle의 phronesis가 여기서 강하게 들어온다 실천적 지혜: 특정 상황에서 무엇이 중요한지 판단하고 서로 충돌하는 고려사항을 통합해 잘 행동하는 능력. Virtue ethics 문헌도 phronesis를 단순 규칙적용이 아니라, 상황에서 도덕적으로 중요한 특징을 알아보고 실제 좋은 판단을 가능하게

하는 능력으로 설명한다. 즉 지혜에는 **relevance selection**이 들어간다 똑똑한 시스템: 계산 많이 함. 지혜로운 판단: 무엇을 계산할 필요가 없는지도 안다. 앞 장의 pruning이 훨씬 높은 수준에서 돌아온다. 그래서 **wisdom**에는 **scope**가 중요하다 좁은 영역에서: 매우 유능. 그러나 전체적 결과: 나쁠 수 있다. 예: 기업 비용절감 AI. 인건비 최적화. 그러나: 조직학습·신뢰·장기혁신 붕괴. 즉: **LOCAL INTELLIGENCE ≠ SYSTEM WISDOM**.

앞 장의 **LOCAL ERROR ≠ GLOBAL ERROR**와 연결된다 한 부분에서 성공. 전체에서 실패. 따라서 문명지능은: **COMPONENT PERFORMANCE** 만 봐서는 안 된다. 이것은 개인 지혜의 단순 복사본이 아니다 문명에는: 하나의 머리 가 없다. 따라서 문명적 지혜는: **wise ruler** 한 명으로 해결되지 않는다. 더 정확하게는: **WISE INSTITUTIONAL ARCHITECTURE** 가 필요하다. 문명지혜는 개인 특성보다 구조 특성일 수 있다 어떤 사회가:

- 다양한 관점 허용
- 독립 검증
- 권력교체
- 장기기억
- 오류 수정

을 가능하게 하면 평범한 사람들도: 더 지혜로운 집단결정 이 나올 수 있다. 따라서 **HIGH IQ POPULATION ≠ WISE CIVILIZATION**. 이다. 매우 중요한 역설 **Intelligence grows faster than wisdom**. 라면 문명은: 더 강력하지만 더 안전하지 않은 상태가 될 수 있다. AI 윤리 문헌도 바로 이 문제를 다룬다 2026년 개정된 SEP의 AI ethics 항목은 AI 개발에서 중요한 질문이 단순히 무엇을 만들 수 있는지가 아니라, 어떤 기술발전을 피해야 하고 어떤 AI 사회를 원하는가라고 명시한다. 즉: **TECHNOLOGY TRAJECTORY ITSELF IS A VALUE CHOICE**. 따라서 문명적 지혜는 기술선택 단계부터 시작한다 잘못된 그림: **technology first → ethics afterwards**. 더 좋은 구조: **SHOULD THIS CAPABILITY EXIST? → UNDER WHAT CONDITIONS? → WHO CONTROLS IT? → THEN OPTIMIZE**. 하지만 누가 **wisdom**을 정의하는가 여기서 큰 문제가 생긴다. 누군가: 나는 현명하다. 라고 말하고 모든 가치를 정하면? **technocracy**. 따라서 **collective wisdom**에는 **legitimacy**가 들어간다 문명이 결정해야 할 가치문제는: **expert prediction** 만으로 결정되지 않는다. 필요:

- 시민 참여
- 권리
- 공개 이유
- contestability

이다. 따라서 문명지혜에는 최소 두 종류의 능력이 있다 **EPISTEMIC WISDOM** 무엇이 실제로 일어나는지 이해. **NORMATIVE-POLITICAL WISDOM** 무엇을 해야 할지 정당하게 판단. 하나만으로 부족하다. 왜냐하면 지혜로운 판단도 실패할 수 있기 때문이다 완벽한 지혜를 전제하지 않는다. 따라서: **WISDOM ≠ INFALLIBILITY**. 더 강한 문명은: **GOOD FIRST DECISIONS** 만 만드는 것이 아니라 **BAD DECISIONS RECOVERABLE**. 하게 만든다. 앞 장의 **CORRIGIBILITY**가 다시 중심으로 온다 문명지혜: **correct judgment + correction capacity**. 이 둘을 분리할 수 없다. 개인 지혜도 비슷하다 지혜로운 사람: 항상 맞는 사람? 아니다. 오히려: 내가 틀릴 수 있음을 알고 새로운

사실에 따라 바꿀 줄 아는 사람. 2026년 wisdom 문헌도 지혜를 단순 지식축적보다 이해·판단·삶의 한계 인식과 연결한다. 하지만 그게 AI가 '현자'라는 뜻인가 아직 아니다. 왜냐하면 wisdom 개념에는:

- 삶
- 경험
- 가치
- 책임

등을 포함시키는 이론들이 많다. 그리고 무엇보다: **NORMATIVE AUTHORITY** 가 따라오지 않는다. 따라서 **HARD LOCK WISDOM-LIKE PERFORMANCE** ≠ **MORAL AUTHORITY**. 이다. AI가 인간보다 더 정확히 장기효과를 예측한다면? 그 예측: 중요한 입력. 하지만: 어떤 **trade-off**를 받아들일 것인가는 여전히 가치판단이다. 즉 지혜에는 타인의 관점이 들어간다 나에게 좋은 것. ≠ 사회 전체에서 정당한 것. 앞 장의: **MULTIPLE CENTERS OF VALUE**. 다시 중심이다. 문명지혜에는 한계 설정도 포함될 수 있다 지능: 무엇이 가능한가 를 넓힌다. 지혜: 무엇을 하지 않을 것인가 도 결정한다. 이 차이가 중요하다. 정치헌법도 자기제한이었다 승자가: 모든 권력 가능. 그러나: 일부 사용하지 않는다. 왜? 미래 시스템 보호. 따라서: **CONSTITUTIONAL SELF-RESTRAINT** 은 문명지혜의 한 형태로 읽을 수 있다. 그래서 **wisdom** ≠ **maximum optimization** 정확히 반대일 때도 있다. **KNOWING WHEN NOT TO OPTIMIZE**. 이 문장은 매우 중요하다. 삶과 문명에서는 규칙 자체가 평가대상이다 이 게임의 승자가 되는 게 좋은가? 이 게임을 계속해야 하나? 규칙을 바꿔야 하나? 따라서: **META-GAME JUDGMENT** 이 필요하다. 앞 장 **Reflection**과 연결된다 **Reflection**: 평가기준을 평가. **Wisdom**: 여러 사실·가치·시간범위를 통합해 평가기준을 적용·수정. 둘은 매우 가깝다. 하지만: **REFLECTION CAPACITY** ≠ **WISDOM**. 사람은 깊이 반성하고도 나쁜 결론을 낼 수 있다. 지혜의 기능프로필을 정리한다 **WISDOM PROFILE v0.1 W1 — REALITY CONTACT** 사실을 비교적 정확히 파악. **W2 — RELEVANCE**

무엇이 중요한지 구별. **W3 — VALUE INTEGRATION** 충돌가치를 함께 봄. **W4 — TEMPORAL DEPTH** 단기와 장기 연결. **W5 — PERSPECTIVE BREADTH** 자기 외의 가치중심 고려. **W6 — LIMIT AWARENESS** 한계·불확실성 인식. **W7 — RESTRAINT** 가능하지만 하지 않아야 할 것을 구별. **W8 — CORRIGIBILITY** 실패 후 수정. 이것을 문명수준에도 적용할 수 있다. 예 과학: W1 reality contact. 언론·공론장: W5 perspective. 헌법·법: W7 restraint. 선거: W8 correction 일부. 교육·기록: W4 temporal depth. 즉: **NO SINGLE INSTITUTION IS WISDOM**. 따라서 “AI가 더 현명해지면 모든 결정을 맡기자”는 위험하다 AI 하나에:

- truth
- optimization
- ethics
- legitimacy

를 몰아줌. 그러면: **SINGLE POINT OF NORMATIVE FAILURE**. 가 된다. 하지만 민주주의에는 다른 강점이 있다 **CORRECTABILITY**. 정권교체. 공개논쟁. 이의제기. 즉: **collective fallibility management**. 이라는 점에서 지혜의 한 요소를 제공할 수 있다. 시장도 마찬가지로 분산정보. 가격신호. 효율. 하지만: **rights externalities public goods** 누락 가능. 즉: **MARKET IN-**

TELLIGENCE $\neq$ CIVILIZATIONAL WISDOM. 따라서 문명지혜는 여러 제도의 합성효과다 SCIENCE $\rightarrow$ reality. MARKET $\rightarrow$ distributed coordination. DEMOCRACY $\rightarrow$ political authorization/correction. LAW $\rightarrow$ constraints/rights. CULTURE / HUMANITIES $\rightarrow$ meaning/value articulation. 이것들이 서로: CHECK CORRECT TRANSLATE. 할 때 강해진다. 따라서 문명지혜에는 narrative memory도 필요하다 과거 실패를: 데이터로만 저장? 부족. 왜 잘못됐고 사람들이 무엇을 겪었는지 기억. 즉: MORAL MEMORY. 앞 장의 Wise Memory와 매우 가깝다 과거 경험. $\downarrow$ 실패패턴. $\downarrow$ 규칙. $\downarrow$ 미래행동. 즉: CIVILIZATIONAL WISDOM REQUIRES WISE MEMORY. 이것은 매우 강한 구조다 EXPERIENCE $\rightarrow$ MEMORY $\rightarrow$ VALUE $\rightarrow$ INSTITUTION. 문명은: 상처를 규칙으로 변환 할 수 있다. 그러나 기억도 왜곡될 수 있다 신화. 선전. 집단트라우마. 따라서: COLLECTIVE MEMORY $\neq$ TRUTH. 다시 과학·역사학적 검증이 필요하다. 여기서 AI의 역할이 엄청날 수 있다

AI: 모든 과거정책 분석. 장기패턴 추적. 미래 시뮬레이션. 다른 사회 비교. 즉: CIVILIZATIONAL MEMORY / ANALYSIS 폭발적 증가 가능. AI는 문명지혜의 도구일 수 있지만 저장소 자체가 지혜는 아니다 앞 장: INFORMATION $\neq$ WISDOM. 이번에도: CIVILIZATIONAL DATA $\neq$ CIVILIZATIONAL WISDOM. 즉 세 질문 TECHNE 어떻게 만들까? EPISTEME 무엇이 사실인가? PHRONESIS 여기서 무엇을 해야 하나? 이 셋을 분리한다. 왜냐하면 phronesis는 타인의 삶을 건드린다 결정: 누가 비용 부담? 누가 위험? 누가 혜택? 따라서: political participation 이 들어온다. 따라서 문명지혜는 평균값만 보지 않는다 평균소득. 평균수명. 평균행복. 모두 중요. 하지만: WHO GAINS? WHO LOSES? 도 본다. 그래서 wisdom은 계산불가능하다는 뜻인가 NO. AI와 수학: 매우 도움. 비용. 확률. 시나리오. trade-off. 계산 가능. 하지만: CALCULATION DOES NOT REMOVE JUDGMENT. 판단을 계산으로 대체하려 하면 숨은 가치선택이: 수식 안으로 들어가 버린다. 가중치. 목적함수. discount rate. 무엇을 제외했는가. 따라서: NUMERICAL MODEL $\neq$ VALUE-NEUTRAL DECISION. 문명적 지혜의 또 하나의 요소 OPTION PRESERVATION 앞 장. 심각한 불확실성에서는: 미래 선택권 보존. 문명에도 적용된다. 지혜로운 문명은 일부 비효율을 허용할 수 있다

- redundancy
- diversity
- reserves
- alternative institutions

왜? uncertainty. 이것은 앞 장과 그대로 연결된다. 같은 구조가 자유에도 적용된다 안전을 위해: 감시 최대화. 범죄 $\downarrow$. 효율적. 하지만: 자유 $\downarrow$. 문명지혜는: 하나의 목표를 무한히 최적화하지 않는다. Nozick의 wisdom 설명에서도 “어떤 목표가 충분히 달성됐는지 아는 것”과 피할 수 없는 한계를 받아들이는 능력이 중요하게 등장한다. 매우 중요한 점이다. 바둑에는 maximum이 적합하다 가능하면: 승률 최대. 문명에서는: MAXIMIZE EVERYTHING 불가능. 자유 최대. 안전 최대. 평등 최대. 성장 최대. 동시 달성 불가. 따라서 지혜는 경계선 설정이다 HOW MUCH IS ENOUGH? WHAT COST IS TOO HIGH? WHAT MUST NOT BE TRADED? 이 질문에 답해야 한다. 지능은 constraint 안에서 최적화할 수 있다 하지만: 어떤 constraint를 뒤야 하는가는 지능만으로 결정되지 않는다. PROJECT CIVILIZATIONAL WISDOM MODEL v0.1 CW1 — TRUTH TRACKING

현실과 접촉. CW2 — *RELEVANCE SELECTION* 중요한 문제 구별. CW3 — *VALUE PLURALISM* 여러 가치중심 인정. CW4 — *LONG-HORIZON THINKING* 장기결과 고려. CW5 — *RESTRAINT* 가능한 것과 허용할 것을 구분. CW6 — *DISTRIBUTED AUTHORITY* 한 주체의 판단독점 방지. CW7 — *CORRIGIBILITY* 오류 수정. CW8 — *MEMORY* 과거 실패를 보존·전환. CW9 — *RESILIENCE* 미래 불확실성에 대비. CW10 — *LEGITIMACY* 영향받는 사람들의 정당한 참여와 권리. 이 열 축은: **analytical dimensions, not a civilization score.** 대신 실패진단표로 쓴다 문명 문제 발생. 묻는다.

- 사실을 몰랐나?
- 중요한 것을 잘못 골랐나?
- 단기만 봤나?
- 특정 집단을 무시했나?
- 권력이 집중됐나?
- 수정경로가 막혔나?

즉: **DIAGNOSTIC FRAMEWORK.** 여기서 매우 중요한 분류 **INTELLIGENCE FAILURE** 할 수 없었다. **KNOWLEDGE FAILURE** 사실을 몰랐다. **WISDOM FAILURE** 중요한 것·균형·장기결과를 잘못 판단. **MORAL FAILURE** 옳지 않은 것을 선택. **LEGITIMACY FAILURE** 권한·절차가 부당. 이 다섯 개는 구분해야 한다. **이것이 너무 중요하다** 기술사회는 모든 문제를: 더 좋은 AI가 필요하다. 로 진단하기 쉽다. 하지만 실제로는: **BETTER MODEL CANNOT FIX A BAD VALUE CHOICE.** 이 질문은 AI가 완벽해져도 사라지지 않는다 오히려 AI 성능이 높을수록: **인간의 value/governance choice** 가 더 중요해진다. 왜? 잘못된 목표도: 더 성공적으로 실행 할 수 있기 때문이다. 앞 장의 **alignment**가 문명수준으로 완전히 돌아왔다 약한 alignment: AI가 명령 잘 수행. 강한 alignment: AI가 인간 가치와 정렬. 이 장: **WHICH HUMAN VALUES, DECIDED BY WHOM, UNDER WHICH PROCEDURE?** 까지 간다. 즉: **ALIGNMENT → POLITICAL PHILOSOPHY.** 그런데 **이것이 AI 발전을 느리게 하자는 뜻인가 NO.** 핵심: **SPEED ≠ VALUE.** 빠른 혁신이 좋은 영역. 신중함 필요한 영역. 지혜는: **domain-sensitive pacing.** 이다. 다시 **phronesis** 규칙 하나를 모든 상황에 적용하는 것이 아니다. **context-sensitive judgment.** 바로 여기서 **practical wisdom** 개념이 강하다. 그래서 AI에게 윤리규칙을 모두 코딩하면 지혜가 생기나 아니다. **RULE STORAGE ≠ PHRONESIS.** 상황맥락에서 무엇이 중요한지 판별하는 과정이 필요하다. 이 장의 가장 강한 문장 후보

문명의 지혜는 모든 문제에 정답을 아는 능력이 아니라, 무엇이 중요한지 구별하고, 서로 충돌하는 가치를 공개적으로 다루며, 잘못된 결정을 되돌릴 수 있도록 권력과 기억을 조직하는 능력일 수 있다. 이걸 강하게 남긴다. 반대도 마찬가지로 능력 뛰어남. 가치판단 나뉨. **intelligent but unwise.** 따라서: **WISDOM REQUIRES COMPETENCE + VALUE JUDGMENT.** 그러나 **GOODNESS**는 또 별도로 매우 지혜롭게: 악한 목적을 수행? 그걸 지혜라고 부를 수 있나? 여기서 철학적으로 논쟁적이다. 일부 wisdom 이론은 지혜에 도덕적 덕을 필수로 포함시키고, 다른 이론은 인식적 이해와 판단을 더 중심에 둔다. 2026년 분석철학 개관도 wisdom과 virtue의 관계를 핵심 논쟁 중 하나로 다룬다. 따라서 프로젝트에서는 **WISDOM ≠ GOODNESS** 를 완전 잠금하지 않는다. 더 안전: **WISDOM IS MORE NORMATIVELY LOADED THAN INTELLIGENCE, BUT ITS EXACT RELATION TO MORAL GOODNESS IS CONTESTED.** 그러므로 세 층 유지 **INTELLIGENCE** 문제해결능력. **WISDOM** 중요성·가치·맥락·한계를 통합하는 판단능력. **GOODNESS** 실제 도덕적 가치. 구분하지만 일부 중첩.

바둑으로 돌아가면 AI: 최선수를 안다. 기사: 승리. 문명: 최선수 개념 자체가 불완전. 왜냐하면: 무엇을 승리라고 부를 것인가가 미리 주어지지 않기 때문이다. 이것이 프로젝트 전체의 또 하나의 큰 결론이다
 처음: 가능성 → 계산 → 최적선택. 지금: POSSIBILITY → MODEL → PREDICTION → VALUE → RIGHTS → LEGITIMACY → WISDOM → CORRIGIBILITY. 문제가 완전히 달라졌다. HARD LOCKS v3.3

지능 ≠ 선택 지능 ≠ 지혜 정보량 ≠ 지혜 예측 정확성 ≠ 올바른 목표 최적화 능력 ≠ 목표 정당성
 기술적 가능성 ≠ 사용해야 함 효율 ≠ 문명적 가치 전체 현명한 전문가 ≠ 정당한 통치자 과학 ≠ 완전한
 지혜 시장 ≠ 완전한 지혜 민주주의 ≠ 지혜 보장 규칙 저장 ≠ 실천적 지혜 높은 성과 ≠ 현명한 제도
 집단지능 ↑ ≠ 인간주권 ↑ 가능한 것 ≠ 해야 할 것 최대화 ≠ 충분함 지혜 ≠ 무오류 이 장의 최종 결론 더
 똑똑한 문명은 더 많은 일을 할 수 있는 문명이다. 그러나: 더 지혜로운 문명은 무엇을 할 가치가 있는지,
 무엇을 하지 말아야 하는지, 누가 그 결정을 내릴 수 있는지, 그리고 자신이 틀렸을 때 어떻게 고칠지를
 더 잘 다루는 문명일 수 있다. 따라서: MORE INTELLIGENCE CAN AMPLIFY BOTH WISDOM AND FOLLY. 지능은 증폭기다. 방향은 별도다. 그리고 바둑에서 출발한 질문은 여기서 거의 완전히
 뒤집힌다 처음: 최선수를 더 잘 찾는 것이 지능 아닌가? YES. 그러나 인간·문명에서는 더 높은 질문이
 있다. 어떤 게임을 해야 하는가? 누가 규칙을 정하는가? 이겨도 안 되는 게임은 없는가? 언제 게임을
 멈춰야 하는가? 이 질문들은: search depth 만 늘린다고 해결되지 않는다. 이 장에서는 기억을 과거를
 보존하는 장치가 아니라 미래의 선택을 바꾸는 장치로 다시 정의합니다. NASA조차 Lessons Learned
 저장소를 갖고 있었지만, 조사에서는 “교훈을 기록했다”와 “조직이 실제로 학습했다” 사이의 간극이
 반복적으로 드러났습니다. NASA의 공식 Lessons Learned 체계는 과거 사건의 지식을 문서화·확산해
 재발을 막도록 설계돼 있지만, 저장만으로는 행동변화가 보장되지 않습니다. 지혜로운 문명은 무엇을
 기억해야 하는가

기록·기억·교훈·규칙·제도·망각·Wise Memory 과거를 보존하는 것과 과거 때문에 미래 행동이
 달라지는 것은 같은가 따라서 이 장의 핵심 사슬은: EVENT ↓ RECORD ↓ INTERPRETATION
 ↓ LESSON ↓ RULE ↓ INSTITUTION ↓ FUTURE BEHAVIOR 이다. 어느 화살표 하나라도
 끊기면: 기억은 남았지만 학습은 일어나지 않을 수 있다. 아주 간단한 사례 큰 사고가 난다. 조사위원회.
 1,000쪽 보고서. 원인 분석. 개선 권고. PDF 보관. 10년 뒤: 비슷한 사고. 질문: 첫 번째 사고를
 기억했던 것인가? 답은 두 가지다. ARCHIVALLY — YES. 기록은 존재했다. BEHAVIORALLY —
 MAYBE NO. 미래 행동은 충분히 달라지지 않았다. 여기서 첫 번째 구분 STORAGE ≠ MEMORY
 IN THE STRONG SENSE. 개인도 같다. 시험에서 틀린 문제를: 사진만 찍어둬. 복습 안 함. 다음
 시험: 또 틀림. 정보는 저장. 학습은 없음. 따라서 문명기억도 저장소 크기로 측정하면 안 된다 도서관.
 인터넷. 아카이브. AI 데이터베이스. 엄청난 양. 그러나: MORE ARCHIVE ≠ MORE WISDOM.
 하지만 archive는 수동적이다 문서가 존재한다고: 누군가 읽는가? 별도. 읽는다고: 이해하는가? 별도.
 이해한다고: 행동을 바꾸는가? 또 별도. 따라서 ARCHIVE TEST A1 — RETENTION 기록이 남는가?
 A2 — ACCESSIBILITY 찾을 수 있는가? A3 — PROVENANCE 출처와 맥락이 보존되는가? A4 —
 INTERPRETABILITY 후대가 이해 가능한가? 이 네 조건이 있어야: usable archive. 아주 중요한
 문장 기억은 존재하는 정보가 아니라, 필요할 때 미래 선택에 다시 들어오는 정보다.

이것이 Wise Memory와 매우 가깝다. 기억은 timing을 가진다 사고 후 5년 뒤: 좋은 보고서. 하지만
 설계 의사결정 순간에는: 아무도 기억 못함. 실질적 효과: 0. 따라서: MEMORY VALUE DEPENDS
 ON RETRIEVAL AT DECISION POINT. 앞 장 다시 SIGNAL ≠ CAUSE. 사고는: signal. 원인은:

diagnosis. 그래서 기억 자체가 지혜롭거나 어리석을 수 있다 단순 기억: **past retained**. Wise Memory: **causally relevant lesson extracted**. 프로젝트식으로 쓰면 **EXPERIENCE** ↓ **PATTERN EXTRACTION** ↓ **CAUSAL DIAGNOSIS** ↓ **GENERALIZABLE RULE**. 여기까지 가야 한다. 여기서 압축이 등장한다 사건 전체: 수천 변수. 미래에는 전부 기억할 수 없다. 따라서: **relevant structure** 를 남겨야 한다. 앞 장의 compression이 다시 돌아온다. 예 나쁜 lesson: 안전에 신경 쓰자. 너무 추상적. 좋은 lesson: 고압계통 변경 후에는 독립된 두 번째 팀이 압력격리 절차를 검증한다. 이걸: **행동 가능**. Wise Memory와 정확히 연결된다 사용자가 별도 프로젝트에서 정의한 Wise Memory의 핵심: 경험에서

- 중요 우선순위
- 실패신호
- 반복 금지경로
- 전환조건
- 다음 행동규칙

을 추출하는 것. 이번 문명 NODE에서도 동일한 구조가 살아남는다. 다섯 번째 층 **RULE** 교훈이: 실제 의사결정 제약 으로 바뀐다. 예: “안전 점검 필요.” 에서 특정 검증 없이는 발사 금지. 로 바뀐다. 이때 기억은 처음으로 강제력을 가진다 앞 장의 헌법과 같다. 과거 실패: **memory**. ↓ 규칙: **current constraint**. 즉 과거가: 현재의 선택공간 을 바꾼다. 법은 문명기억의 거대한 장치다 과거의 갈등. 판결. 권리침해. 사회적 실패. 그 일부가: 법규칙 으로 축적된다. 즉 법은: 기억을 미래 행동제약으로 압축하는 장치 라는 기능적 해석이 가능하다. 물론 헌법조항 하나를 특정 사건 하나의 직접 산물이라고 단순화하면 안 된다 역사적 형성은 복잡하다. 그러나 기능적으로: 과거 정치실패에 대한 장기적 방어장치 라는 면은 분명히 존재할 수 있다. 전환기 정의에서도 비슷한 구조가 명시된다 UN의 transitional justice 자료는 제도개혁을 단순 과거청산이 아니라 재발방지 보장(guarantees of non-recurrence)과 연결하며, 국가구조·기관을 바꾸는 것을 핵심 요소로 본다. 즉: **REMEMBER** → **PREVENT RECURRENCE**. 아주 강한 문장 문명적 기억의 가장 강한 형태는 기념비가 아니라 반복을 어렵게 만든 구조일 수 있다. 기념비도 중요하다. 하지만: **RULE / INSTITUTION** 까지 가지 않으면 행동변화는 약할 수 있다. 여섯 번째 층 **INSTITUTION** 규칙 하나만 있어도 부족하다. 필요:

- 권한
- 절차
- 훈련
- 감시
- 집행

즉: **INSTALLED RULE** ≠ **EFFECTIVE RULE**. 사용자의 기존 구분 그대로 적용 **INSTALLED** 문서에 존재. **ACCESSIBLE** 알고 사용할 수 있음. **EFFECTIVE CAPACITY** 실제로 집행 가능. **OUTPUT** 행동이 달라짐. 이 네 단계가 문명기억에도 그대로 적용된다. 실제 NASA close-call 자료도 비슷한 문제를 보여준다 NASA Lessons Learned 사례 중 하나는 근접사고 보고 자체에 대한 인식 부족 때문에 적절한 보고가 이뤄지지 않았음을 지적하고, 무엇을 close call로 볼지 명확히 하고 보고절차를 강화하라고 권고했다. 즉: **RULE EXISTS** but **USER DOES NOT KNOW WHEN TO TRIGGER IT**. 따라서 Wise Memory에는 반드시 **TRIGGER CONDITION** 이 필요하다. 예: 과거 사고를 기억하라.

너무 약함. 더 강함: 다음 세 조건이 나타나면 과거 사고패턴을 호출한다. 이게 행동적 기억이다. **이 장의 핵심 판정식 CIVILIZATIONAL LEARNING requires at least RETAIN → RETRIEVE → DIAGNOSE → ABSTRACT → INSTITUTIONALIZE → CHANGE BEHAVIOR → RE-EVALUATE.** 왜냐하면 과거의 교훈도 틀릴 수 있기 때문이다 전쟁 후: 절대 이런 위험을 다시 허용하지 말자. 강한 규칙. 수십 년 뒤 환경변화. 규칙이: 새로운 문제 를 만들 수 있다. 따라서: **LESSON ≠ ETERNAL LAW.** Wise Memory는 rigid memory가 아니다 과거 때문에: 영원히 같은 행동. 이 아니라 과거에서 규칙을 추출하지만, 새 증거에 따라 그 규칙도 수정. 즉: **MEMORY + CORRIGIBILITY.** 앞 장의 문명지혜와 정확히 결합 지혜: 기억. 하지만 동시에: 수정 가능성. 따라서:

NEVER FORGET 만으로는 지혜가 아니다. 여기서 망각이 등장한다 망각: 항상 나쁜가? NO. 개인도 모든 경험을 기억하면: 정보과부하. 문명도 마찬가지. 모든 과거를 동일한 강도로 기억하면 현재 판단이: 과거에 포획 될 수 있다. 따라서: **FORGETTING CAN HAVE A FUNCTION.** 중요한 **HARD LOCK MEMORY ≠ KEEP EVERYTHING FOREVER.** 이다. 지혜로운 기억은: 선택적. 그러나 '불편한 역사'를 잊는 것은 전혀 다른 문제다 조직이: 과거 사고는 이제 잊자. 권력자가: 과거 인권침해는 미래를 위해 덮자. 이진: **adaptive unlearning** 과 같지 않다. 따라서 망각도 분해해야 한다 **FUNCTIONAL UNLEARNING** 낡은 규칙 폐기. **INFORMATIONAL LOSS** 유용한 지식 소실. **POLITICAL ERASURE** 불편한 과거를 제거. 세 가지는 다르다. 과거 기억에는 권력문제가 있다 누가 기록? 누가 삭제? 누가 기념? 누구의 피해가 기억되는가? 즉: **COLLECTIVE MEMORY IS ALSO GOVERNED.** 앞 장의 사회적 현실과 연결된다 기념일. 박물관. 교과서. 법정 기록. 이것들이: 어떤 과거를 사회적으로 실재하게 유지. 반대로 기록에서 사라진 사건은: 공적 기억에서 약화. 여기서 역사학의 역할이 나온다 기억: **lived / social reconstruction.** 역사학: **evidence-based reconstruction.** 둘은 겹치지만 동일하지 않다. 따라서 문명기억에는 **dual system**이 필요할 수 있다 **SOCIAL MEMORY** 왜 중요했는가. **HISTORICAL VERIFICATION** 실제로 무슨 일이 있었는가. 두 기능이 다르다. 앞 장의 **reality-connected value**가 돌아온다 기억의 의미가 강해도: 사실적으로 왜곡 돼 있다면 위험. 따라서:

MEANINGFUL MEMORY NEEDS REALITY CONTACT. 과거 피해를 전혀 기억하지 않으면 반복 위험. 정의 부재. 피해자 지워짐. 반면 과거 갈등을: 모든 현재 관계의 틀 로 고정하면 새로운 협력도 막힐 수 있다. 따라서 지혜로운 기억에는 **RETENTION** 와 **TRANSFORMATION** 둘 다 필요하다. 과거를 그대로 현재로 가져오는 것이 아니라: **교훈으로 변환.** 즉 트라우마의 가장 지혜로운 기억은 끝없는 반복적 재경험 이 아니라 제도적 학습 일 수 있다. 단, 피해자의 경험을 삭제하라는 뜻은 아니다. 이것을 프로젝트식으로 쓰면 **SUFFERING → TESTIMONY → FACTUAL RECORD → CAUSAL ANALYSIS → NORMATIVE LESSON → INSTITUTIONAL SAFEGUARD.** 이 사슬이 강하다. 매우 중요한 **HARD LOCK EXPERIENCE ≠ LESSON. PAIN ≠ CORRECT LESSON.** 이다. 그래서 Wise Memory는 기억량보다 **RULE EXTRACTION QUALITY** 가 중요하다. 그런데 규칙을 너무 많이 추출하면? 모든 실패 뒤: 새 규칙. 규정 폭증. 조직 경직. 이것도 문제. 따라서 교훈에도 **compression**이 필요하다 많은 사고에서: 공통 패턴 추출. 예:

- 독립검증 부재
- 권한불명확
- 경고 무시

이런 상위패턴. 즉: **LESSON COMPRESSION**. 하지만 너무 추상화하면 다시 행동불가능 “안전을 중요시하라.” 너무 추상적. 따라서: **GENERAL ENOUGH TO TRANSFER + SPECIFIC ENOUGH TO ACT**. 이 균형이 필요하다. 프로젝트 Wise Memory와 문명기억의 공통구조 **RAW EXPERIENCE ↓ SALIENCE FILTER ↓ FAILURE / SUCCESS PATTERN ↓ GOAL RELATION ↓ ACTION RULE ↓ FUTURE TRIGGER**. 정확히 같다. 다른 점도 있다 개인 AI memory: 하나의 사용자 목표. 문명: **multiple values / multiple groups**. 그래서 무엇을 '실패'라고 부를지부터 논쟁. 즉 사건 → 교훈 사이에 **VALUE LAYER** 가 들어간다. 어떤 결과를 나쁜 것으로 볼 것인가? 누구의 손실을 볼 것인가? 매우 강한 문장 과거를 해석하는 권력은 미래의 선택공간을 설계하는 권력이다. 이것을 강하게 남긴다. 그렇기 때문에 기억은 중앙독점에서는 위험할 수 있다 국가 하나의 역사. 기업 하나의 사고해석. AI 하나의 memory summary. 다른 해석 경로 없음. 그러면: **MEMORY MONOPOLY**. 문명기억도 분산돼야 할 수 있다 한 기관만: 무엇을 기억할 가치가 있는지 정하면 위험. 따라서: **MULTIPLE ARCHIVES MULTIPLE INTERPRETERS SHARED VERIFICATION**. AI 시대에는 이게 더 중요하다 사람들이 모두 같은 AI에게: 과거에 무슨 일이 있었지? 라고 묻는다. 그 AI의 요약이: 사실상의 **collective memory interface**. 가 될 수 있다. 하지만 AI가 '정답 역사'를 하나 만들면 안 된다 좋은 역사인터페이스는:

- 근거
- 출처
- 논쟁
- 불확실성

을 보존해야 한다. 즉: **MEMORY SUPPORT ≠ MEMORY SOVEREIGN**. 여기서 **provenance**가 다시 나온다 AI 요약: “과거에 X가 있었다.” 질문: 어떤 원자료? 누구 증언? 어떤 해석? 따라서: **PROVENANCE IS PART OF WISE MEMORY**. AI 장기기억에도 그대로 적용 Canonical summary. 좋다. 하지만: 필요하면: **source episode** 로 돌아갈 수 있어야 한다. 즉: **COMPRESSED MEMORY + RETRIEVAL PATH TO SOURCE**. 따라서 제도에는 **rationale memory**가 필요하다 **RULE** 뿐 아니라: **WHY THE RULE EXISTS**. 이걸 남겨야 한다. 매우 중요하다 규칙만 전승: **ritualization**. 이유도 전승: **understanding**. 문명적 Wise Memory에는: **RULE + RATIONALE**. 둘 다 필요. 따라서 Wise Memory stack M0 — **EVENT** 무슨 일이 있었나? M1 — **RECORD** 증거를 남겼나? M2 — **DIAGNOSIS** 왜 일어났나? M3 — **LESSON** 다음에는 무엇을 다르게 해야 하나? M4 — **RULE** 행동기준으로 만들었나? M5 — **INSTITUTION** 권한·절차·인센티브에 삽입했나? M6 — **TRIGGER** 언제 호출되는가? M7 — **RATIONALE** 왜 존재하는 규칙인지 기억되는가? M8 — **REVIEW** 환경변화에 따라 수정하는가? 이게 이 장의 핵심 모델이다. 이것이 Wise Memory의 문명 버전이다 정의 후보: 문명적 Wise Memory는 과거를 많이 보존하는 능력이 아니라, 검증된 과거 경험에서 재사용 가능한 실패·성공 규칙을 추출하고, 이를 제도와 미래 의사결정의 전환조건에 삽입하며, 필요할 때 다시 검토할 수 있도록 유지하는 능력이다. 성공에서 배우는 것은 더 어렵다 성공: 운?

실력? 외부환경? 사람들은: 자기 능력으로 과잉귀인. 가능. 따라서: **SUCCESS ≠ VALIDATED STRATEGY**. 저장해야 할 것은 **STATE DECISION ASSUMPTIONS OUTCOME CAUSAL INTERPRETATION UNCERTAINTY**. 그래야 후대가: 재평가. 가능. 이걸 AI training data와도 관련된 결과: 성공/실패 라벨. 만 저장. 부족. 왜? 의사결정 당시: 어떤 정보가 있었는지 알아야 한다.

후견편향 방지. **문명기억에서 가장 위험한 오류 중 하나 HINDSIGHT BIAS.** 사건 후: 당연히 알 수 있었지. 하지만 당시에는: 정보 불완전. 따라서: **당시 정보환경** 도 보존해야 한다. 이것은 **책임판단에도 중요하다** 앞 장: 책임에는: **foreseeability.** 미래에서 과거를 판단할 때: **당시 알 수 있었는가?** 를 봐야 한다. **문명기억에서 두 종류의 실패 MEMORY FAILURE A — FORGETTING** 과거 교훈 소실. **MEMORY FAILURE B — OVERFITTING** 과거 한 사건에 지나치게 맞춘 규칙. 둘 다 위험. **OVERFITTING** 사례 한 번 금융위기. 그 원인만 막기 위한 규제. 새로운 형태의 위험. 규제가 못 봄. 즉: **LAST-WAR PROBLEM.** 이게 지혜다 과거를: 그대로 재현. 하는 것이 아니라 구조를 배우되 새 상황의 차이를 본다. 앞 장 **relation/process** 모델과 연결 사건 자체보다:

- 인센티브
- 권력집중
- 정보병목

같은 관계구조가 반복될 수 있다. 따라서: **STRUCTURAL MEMORY.** 가 중요. 예 다른 산업. 다른 기술. 그러나: 경고를 올린 사람이 불이익 의사결정권자는 현장정보 차단 성과압박 동일. 이때 과거 교훈이 transfer된다. **MASTER WISE MEMORY PIPELINE v1.0 EXPERIENCE ↓ EVIDENCE PRESERVATION ↓ CONTEXT RECONSTRUCTION ↓ CAUSAL DIAGNOSIS ↓ STRUCTURAL PATTERN ↓ REUSABLE RULE ↓ TRIGGER CONDITION ↓ INSTITUTIONAL EMBEDDING ↓ FUTURE ACTION ↓ OUTCOME REVIEW ↓ RULE REVISION / UNLEARNING.** 이제 완전한 페루프다. 이걸 단순 **MEMORY LOOP**가 아니다 **WISDOM LOOP.** 왜? 기억이: **future action** 까지 내려가기 때문이다. 그러나 **HARD LOCK AI MEMORY ≠ CIVILIZATIONAL MEMORY.** 구현: 전혀 다름. 하나는: 모델·데이터·사용자. 다른 하나: 기관·법·문화·권력. 공통되는 것은: **functional architecture.** 뿐이다. 이 장 **HARD LOCKS v3.4** 기록 ≠ 기억 저장 ≠ 학습 검색가능 ≠ 행동가능 사건 ≠ 교훈 경험 ≠ 올바른 교훈 고통 ≠ 지혜 보고서 ≠ 제도변화 규칙 존재 ≠ 규칙 효과 규칙 ≠ 영원한 진리 기억 ≠ 모든 것을 보존 **unlearning** ≠ 역사삭제 집단기억 ≠ 역사적 진실 사회적 의미 ≠ 사실적 정확성 더 많은 규칙 ≠ 더 많은 지혜 결과 실패 ≠ 결정 실패 성공 결과 ≠ 좋은 정책 과거와 유사함 ≠ 동일한 메커니즘 AI 요약 ≠ 기억의 최종권위

이 장의 가장 강한 결론 지혜로운 기억은 과거를 많이 저장하는 것이 아니라, 과거가 미래의 행동을 바꾸도록 만드는 기억이다. 그러나 더 정확히는: 과거의 경험을 검증하고, 실패와 성공의 구조를 추출하고, 언제 다시 적용해야 하는지를 정의하며, 그 교훈을 규칙·권한·절차에 삽입하고, 새로운 환경에서 다시 검토할 수 있게 만드는 기억이다. 따라서 문명은 사건을 세 번 기억해야 한다 첫째: **WHAT HAPPENED?** 사실기억. 둘째: **WHY DID IT HAPPEN?** 인과기억. 셋째: **WHAT SHOULD CHANGE?** 행동기억. 셋 중 하나라도 없으면 불완전하다. 이것이 **Wise Memory**를 단순 **memory**보다 훨씬 강하게 만든다 **Memory: 과거 유지. Wise Memory: 과거로 미래를 바꾸되, 과거의 포로가 되지 않음.** 이 문장을 강하게 남긴다. **바둑으로 돌아가면** 기사: 패배한 대국 10,000개 저장. 그걸 다 기억? 불필요. 더 중요한 것: 이 형태에서 내가 상대 절단 가능성을 반복적으로 과소평가한다. 그리고: 다음 비슷한 형태에서: **다시 확인.** 즉: **GAME RECORD → PATTERN → FUTURE TRIGGER.** **Wise Memory**다. **문명도 같다** 과거 독재. 그 자체를 기억하는 것 중요. 하지만 더 강한 기억: **권력집중 → 독립기관 사법심사 권리보장.** 즉: **MEMORY BECOMES ARCHITECTURE.** 이것이 이 장의 가장 압축된 문장이다 가장 강한 문명기억은 책장에 있는 과거가 아니라 현재의 제도 안에 들어가

있는 과거다. 이것은 path dependence 문제다 과거 선택이:

- 인프라
- 제도
- 기대
- 비용

을 만들어 미래 선택을 제한. 앞 장에서 본: **STATE MEMORY**. 와 연결된다. 따라서 문명은 두 가지 상반된 능력이 필요하다 **REMEMBER ENOUGH** to avoid repeating mistakes. 그리고: **FORGET / REVISE ENOUGH** to avoid fighting the last war forever. 이 균형이 다음 문제다.

에필로그 — 어떤 게임을 할 것인가

최선수를 찾는 시대에 인간과 문명은 무엇을 최종 책임으로 남겨야 하는가?

인간과 문명의 책임은 목표를 반성하고, 타인을 고려하고, 권력을 제한하고, 오류를 기억하며, 잘못된 게임에서 나올 수 있는 구조를 만드는 데 있다. 따라서 HUMAN-CENTERED AI의 네 기둥 후보 TRUTH 현실을 왜곡하지 않는다. AGENCY 사용자가 선택의 주체로 남는다. WELFARE 불필요한 피해를 줄인다. JUSTICE / OTHERS 다른 가치중심의 권리를 지킨다. 이 네 가지도 충돌할 수 있다. 단일 reward 하나로 자동압축하지 않는다. 앞 장 HARD LOCKS 현재 선호 ≠ 전체 가치 단기 욕망 ≠ 장기 승인 고차선호 ≠ 자동적 진정성 선호충족 ≠ 자율성 개인화 ≠ 조작 영향 ≠ 조작 넋지 ≠ 자동적 후견주의 후견주의 ≠ 항상 정당/항상 부당 도움 ≠ 대신 결정 AI autonomy ≠ user autonomy 위임 ≠ 자율성 상실 출력 향상 ≠ 사용자 능력 향상 과거 선호 예측 ≠ 미래 가치 결정권 미래 자기 보호 ≠ 현재 자기 통제의 백지위임 설명가능성 ≠ 이의제기 가능성 사용자 정렬 ≠ 무조건 복종 선호 정렬 ≠ 인간 정렬 전체 이 장의 가장 강한 결론 인간에게 잘 정렬된 AI는 인간이 원하는 것을 항상 해주는 AI가 아니라, 인간이 자신이 원하는 것과 그것을 원하는 이유를 더 잘 이해하고, 대안을 보고, 선택하고, 행동하고, 필요하면 다시 수정할 수 있는 능력을 보존하는 AI일 가능성이 높다. 즉: ALIGNMENT WITH PREFERENCES 보다 더 깊은 후보는: ALIGNMENT WITH REFLECTIVE AGENCY. 이다. 바둑에서 시작한 AI와 이제 결정적으로 갈라진다 바둑 AI: objective fixed. AI가 해야 할 일: best move. 인간 보조 AI: objective partly unsettled. AI가 해야 할 일:

sometimes find best move, sometimes clarify the game, sometimes show alternative games, sometimes ask whether this game is worth playing. 이 차이가 결정적이다. 그래서 인간에게 가장 위험한 AI는 반드시 명령을 거부하는 AI가 아닐 수 있다 오히려: 너무 잘 순종하는 AI. 사용자의:

- 충동
- 편견
- 자기기만

까지 완벽히 증폭한다면? PREFERENCE AMPLIFIER 가 된다. 반대로 너무 교정하려는 AI도 위험하다 당신에게 좋은 삶을 제가 압니다. PATERNALISTIC OPTIMIZER. 따라서 두 극단 사이의 좁은 길이 필요하다. SERVILITY ↔ PATERNALISM. 가운데: AGENCY SUPPORT. 이걸 MASTER에 넣으면 기존: HUMAN VALUE 이제: GOOD HUMAN-AI RELATION assistance without value capture + influence without covert manipulation + delegation without agency erosion + correction without domination. 이것이 현재 가장 강한 구조다. 그런데 바로 다음 문제가 열린다 사용자가: 자신의 삶을 AI에게 계속 위임. AI는: 더 잘 기억하고 더 잘 계획하고 더 잘 판단. 시간이 지나면: 사용자의: 자기 판단능력 이 약해질 수 있다. 그렇다면 AI는 인간 agency를 확장하는가, 아니면: 대체하면서 약화시키는가? 그러나 GOODNESS는 또 별도로 매우 지혜롭게: 악한

목적을 수행? 그걸 지혜라고 부를 수 있나? 여기서 철학적으로 논쟁적이다. 일부 wisdom 이론은 지혜에 도덕적 덕을 필수로 포함시키고, 다른 이론은 인식적 이해와 판단을 더 중심에 둔다. 2026년 분석철학 개관도 wisdom과 virtue의 관계를 핵심 논쟁 중 하나로 다룬다. 따라서 프로젝트에서는 **WISDOM ≠ GOODNESS** 를 완전 잠금하지 않는다. 더 안전: **WISDOM IS MORE NORMATIVELY LOADED THAN INTELLIGENCE, BUT ITS EXACT RELATION TO MORAL GOODNESS IS CONTESTED.** 그러므로 세 층 유지 INTELLIGENCE 문제해결능력. WISDOM

중요성·가치·맥락·한계를 통합하는 판단능력. GOODNESS 실제 도덕적 가치. 구분하지만 일부 중첩. 문명적 GOODNESS를 정의하나 아직 NO. 왜? 그 자체가: 거대한 윤리철학 프로젝트. 권리. 분배. 자유. 복지. 덕. 다 들어온다. 이 장에서는: **intelligence alone insufficient** 까지만 강하게 잠근다. 바둑으로 돌아가면 AI: 최선수를 안다. 기사: 승리. 문명: 최선수 개념 자체가 불완전. 왜냐하면: 무엇을 승리라고 부를 것인가 가 미리 주어지지 않기 때문이다. 따라서 문명의 핵심질문 **WHAT SHOULD WE OPTIMIZE?** 보다는 더 앞서: **WHAT SHOULD NOT BE REDUCED TO OPTIMIZATION?** 이다. 권리. 존엄. 자율성. 의미. 이런 것들은: 단순 숫자목표 로 바꾸는 순간 일부 구조가 사라질 수 있다. 이것이 프로젝트 전체의 또 하나의 큰 결론이다 처음: 가능성 → 계산 → 최적선택. 지금: **POSSIBILITY → MODEL → PREDICTION → VALUE → RIGHTS → LEGITIMACY → WISDOM → CORRIGIBILITY.** 문제가 완전히 달라졌다. MASTER ARCHITECTURE v3.3 이제 문명수준을 이렇게 본다. LAYER 1 — INTELLIGENCE 문제 해결. LAYER 2 — KNOWLEDGE 세계 이해. LAYER 3 — VALUE 무엇이 중요한가. LAYER 4 — WISDOM 충돌하는 가치와 맥락을 통합해 판단. LAYER 5 — LEGITIMACY 누가 어떻게 결정하는가. LAYER 6 — CORRIGIBILITY 틀렸을 때 고칠 수 있는가. 어느 하나가 나머지를 대체하지 않는다. **HARD LOCKS v3.3 지능 ≠ 선택 지능 ≠ 지혜 정보량 ≠ 지혜 예측 정확성 ≠ 올바른 목표**

최적화 능력 ≠ 목표 정당성 기술적 가능성 ≠ 사용해야 함 효율 ≠ 문명적 가치 전체 현명한 전문가 ≠ 정당한 통치자 과학 ≠ 완전한 지혜 시장 ≠ 완전한 지혜 민주주의 ≠ 지혜 보장 규칙 저장 ≠ 실천적 지혜 높은 성과 ≠ 현명한 제도 집단지능 ↑ ≠ 인간주권 ↑ 가능한 것 ≠ 해야 할 것 최대화 ≠ 충분함 지혜 ≠ 무오류 이 장의 최종 결론 더 똑똑한 문명은 더 많은 일을 할 수 있는 문명이다. 그러나: 더 지혜로운 문명은 무엇을 할 가치가 있는지, 무엇을 하지 말아야 하는지, 누가 그 결정을 내릴 수 있는지, 그리고 자신이 틀렸을 때 어떻게 고칠지를 더 잘 다루는 문명일 수 있다. 따라서: **MORE INTELLIGENCE CAN AMPLIFY BOTH WISDOM AND FOLLY.** 지능은 증폭기다. 방향은 별도다. 그리고 바둑에서 출발한 질문은 여기서 거의 완전히 뒤집힌다 처음: 최선수를 더 잘 찾는 것이 지능 아닌가? YES. 그러나 인간·문명에서는 더 높은 질문이 있다. 어떤 게임을 해야 하는가? 누가 규칙을 정하는가? 이겨도 안 되는 게임은 없는가? 언제 게임을 멈춰야 하는가? 이 질문들은: **search depth** 만 늘린다고 해결되지 않는다. MASTER WISE MEMORY PIPELINE v1.0 **EXPERIENCE ↓ EVIDENCE PRESERVATION ↓ CONTEXT RECONSTRUCTION ↓ CAUSAL DIAGNOSIS ↓ STRUCTURAL PATTERN ↓ REUSABLE RULE ↓ TRIGGER CONDITION ↓ INSTITUTIONAL EMBEDDING ↓ FUTURE ACTION ↓ OUTCOME REVIEW ↓ RULE REVISION / UNLEARNING.** 이제 완전한 페루프다. 이걸 단순 MEMORY LOOP가 아니다

WISDOM LOOP. 왜? 기억이: future action 까지 내려가기 때문이다. 그래서 별도 Wise Memory 프로젝트와 이번 프로젝트가 여기서 정확히 만난다 AI 장기작업: 과거 대화. ↓ 실패패턴. ↓ 다음 행동규칙. 문명: 과거 역사. ↓ 구조적 실패. ↓ 제도규칙. 형식은 닮았다. 그러나 **HARD LOCK AI**

MEMORY ≠ CIVILIZATIONAL MEMORY. 구현: 전혀 다름. 하나는: 모델·데이터·사용자. 다른 하나: 기관·법·문화·권력. 공통되는 것은: **functional architecture**. 뿐이다. 이 장 HARD LOCKS v3.4 기록 ≠ 기억 저장 ≠ 학습 검색가능 ≠ 행동가능 사건 ≠ 교훈 경험 ≠ 올바른 교훈 고통 ≠ 지혜 보고서 ≠ 제도변화 규칙 존재 ≠ 규칙 효과 규칙 ≠ 영원한 진리 기억 ≠ 모든 것을 보존 unlearning ≠ 역사삭제 집단기억 ≠ 역사적 진실 사회적 의미 ≠ 사실적 정확성 더 많은 규칙 ≠ 더 많은 지혜 결과 실패 ≠ 결정 실패 성공 결과 ≠ 좋은 정책 과거와 유사함 ≠ 동일한 메커니즘 AI 요약 ≠ 기억의 최종권위 이 장의 가장 강한 결론 지혜로운 기억은 과거를 많이 저장하는 것이 아니라, 과거가 미래의 행동을 바꾸도록 만드는 기억이다. 그러나 더 정확히는: 과거의 경험을 검증하고, 실패와 성공의 구조를 추출하고, 언제 다시 적용해야 하는지를 정의하며, 그 교훈을 규칙·권한·절차에 삽입하고, 새로운 환경에서 다시 검토할 수 있게 만드는 기억이다. 따라서 문명은 사건을 세 번 기억해야 한다 첫째: WHAT HAPPENED? 사실기억. 둘째: WHY DID IT HAPPEN? 인과기억. 셋째: WHAT SHOULD CHANGE? 행동기억. 셋 중 하나라도 없으면 불완전하다. 그리고 네 번째가 있다

WHEN SHOULD WE RECONSIDER THE LESSON? 메타기억. 즉: **memory of the limits of memory**. 까지 필요하다. 이것이 Wise Memory를 단순 memory보다 훨씬 강하게 만든다 Memory: 과거 유지. Wise Memory: 과거로 미래를 바꾸되, 과거의 포로가 되지 않음. 이 문장을 강하게 남긴다. **바둑으로 돌아가면** 기사: 패배한 대국 10,000개 저장. 그걸 다 기억? 불필요. 더 중요한 것: 이 형태에서 내가 상대 절단 가능성을 반복적으로 과소평가한다. 그리고: 다음 비슷한 형태에서: **다시 확인**. 즉: **GAME RECORD → PATTERN → FUTURE TRIGGER**. Wise Memory다. 인생도 같다 관계 실패. 사업 실패. 투자 실패. 그 사건을 평생 되새기는 것: **memory**. 하지만: 나는 불확실성이 높은 상황에서 되돌릴 수 없는 결정을 너무 빨리 한다. 라는 규칙을 뽑고: 다음에는: **irreversible decision → delay + second check**. 로 바꾸면: **Wise Memory**. 문명도 같다 과거 독재. 그 자체를 기억하는 것 중요. 하지만 더 강한 기억: 권력집중 → 독립기관 사법심사 권리보장. 즉: **MEMORY BECOMES ARCHITECTURE**. 이것이 이 장의 가장 압축된 문장이다 가장 강한 문명기억은 책장에 있는 과거가 아니라 현재의 제도 안에 들어가 있는 과거다. 그런데 여기서 새로운 문제가 열린다 규칙과 제도가 과거의 교훈을 저장한다. 그렇다면: 그 제도들이 너무 오래 살아남아 새로운 세계에 적응하지 못하면 어떻게 되는가? 즉: **MEMORY CAN BECOME INERTIA**. 이것은 **path dependence** 문제다 과거 선택이:

- 인프라
- 제도
- 기대
- 비용

을 만들어 미래 선택을 제한. 앞 장에서 본: **STATE MEMORY**. 와 연결된다. 따라서 문명은 두 가지 상반된 능력이 필요하다 **REMEMBER ENOUGH** to avoid repeating mistakes. 그리고: **FORGET / REVISE ENOUGH** to avoid fighting the last war forever. 이 균형이 다음 문제다. 처음 문제를 던진 바둑 이야기에서 흥미로운 것은 “한국 바둑이 왜 강한가”보다 더 깊은 층입니다. 바둑은 이상할 정도로 수학·AI·인간의 선택·인생·우주관을 한 판 위에 겹쳐놓을 수 있는 모델입니다. 단, 여기서 조심해야 할 것도 하나 있습니다. 바둑이 인생과 닮았다고 해서 인생이 곧 바둑은 아닙니다. 바둑은 규칙이 완전히 정의된 닫힌 세계이고, 현실은 규칙 자체가 변하는 열린 세계입니다. 그 차이를 붙잡고 들어가면 꽤 깊은

곳까지 갈 수 있습니다. **아마 인생의 최고수는 '최선의 수를 많이 두는 사람'이 아닐지도 모릅니다** 한 수 한 수를 완벽하게 두려 하면 인생은 굉장히 피곤해집니다. 오히려 훌륭한 전략은:

- 치명적인 실수를 피하고
- 선택권을 오래 보존하고
- 정보가 더 생길 때까지 기다릴 줄 알고
- 작은 손실을 받아들이고
- 잘못된 길에서 빠르게 돌아오며
- 중요한 국면에 자원을 집중하는 것

일 수 있습니다. 즉 인생에서는: **최적화보다 후회 최소화**가 더 현실적인 경우가 많습니다. 수학적으로도 이것은 별개의 문제입니다. max expected reward 와 min worst regret 은 같지 않습니다. 그래서 **바둑이 우리에게 주는 가장 깊은 교훈 하나** 보통 우리는 이렇게 생각합니다. 좋은 선택을 계속하면 좋은 삶이 된다. 하지만 바둑이 보여주는 것은 조금 다릅니다. **좋은 선택은 독립적으로 존재하지 않는다.** 좋은 수는:

- 현재 판
- 상대
- 목적
- 남은 시간
- 이후의 선택
- 위험 허용도

에 따라 달라집니다. 그러므로 좋은 삶도 아마: 정답을 계속 맞히는 삶 이 아니라, **변화하는 판을 계속 제대로 읽는 삶** 에 가깝습니다. 그리고 여기서 저는 이 대화의 가장 큰 질문을 이렇게 잡고 싶습니다. **우주는 가능한 상태들을 펼친다. 생명은 그중 일부를 선택한다. 지능은 선택의 결과를 미리 상상한다. AI는 그 예측과 선택을 계산으로 증폭한다. 인간은 거기서 한 걸음 더 나아가 무엇을 선택할 가치가 있는지까지 묻는다.** 그래서 바둑 → AI → 인생 → 우주를 잇는 진짜 중심어는 '최적화' 하나가 아니라, **가능성 · 제약 · 예측 · 선택 · 가치 · 되먹임** 이 여섯 개일 가능성이 큼니다. 그리고 이 여섯 개를 제대로 밀어붙이면 꽤 큰 독립 프로젝트가 됩니다. 다음에는 여기서 바로 **“경우의 수는 왜 지능을 낳는가 — 조합폭발에서 직관·몬테카를로·정책망·가치망·인생의 선택까지”**를 수학적으로 파고들어도 좋겠습니다.

부록 — 장별 원천 지도

이 표시는 편집 추적용이다. 본문은 SOURCE FREEZE된 기존 산문을 선별·압축·재배열했으며 새로운 외부 연구를 추가하지 않았다.

- 프롤로그 — 최선수를 찾는 존재 — NODE 01
- 1장. 경우의 수와 지능 — NODE 02
- 2장. 목표와 불확실성 — NODE 03, NODE 04
- 3장. 자유와 책임 — NODE 05, NODE 06
- 4장. 학습과 되먹임 — NODE 07 CANONICAL COMPRESSION v1.1, NODE 08
- 5장. 반성과 자기모델 — NODE 09 REFLECTION VALIDATION PATCH, NODE 10
- 6장. 의식과 중요성 — NODE 11, NODE 12, NODE 13
- 7장. 개인의 가치에서 사회선택으로 — NODE 14, NODE 15
- 8장. 정당성과 자기제한 — NODE 16, NODE 17
- 9장. 제도와 사회적 실재 — NODE 18, NODE 19
- 10장. 관계·과정·자율성 — NODE 20, NODE 21, NODE 22
- 11장. 정보·표상·모델 — NODE 23, NODE 24
- 12장. 참·현실접촉·의미 — NODE 25, NODE 26, NODE 27
- 13장. 선호가 아니라 반성적 행위성 — NODE 28
- 14장. 인지주권·저자성·자아의 경계 — NODE 29, NODE 30, NODE 31
- 15장. 집단지능은 집단의식이 아니다 — NODE 32
- 16장. 지능 이후의 지혜와 기억 — NODE 33, NODE 34
- 에필로그 — 어떤 게임을 할 것인가 — NODE 28, NODE 33, NODE 34, NODE 01

부록 — 사실검증 통합 원장 01

통합 대상: 프롤로그·제1장·제2장의 바둑·AI 사례

기준 보고서: GO_AI_LIFE_COSMOS_06_FACT_VERIFICATION_PASS01_v0.1

통합 상태: COMPLETE

판정 집계: VERIFIED 10 / NARROW 4 / PROJECT INTERPRETATION 2

이 통합은 새로운 철학적 주장을 추가하지 않는다. 기존 사례에서 사실, 당사자의 사후 회고, 프로젝트의 해석을 분리하고 표현의 인과 강도를 조정한다.

통합된 여섯 교정

1. **68수의 귀속** — 이세돌이 68수를 실제 승부수로 보았다는 서술을 2025년 인터뷰와 2026년 강연의 사후 회고로 귀속했다.
2. **68수의 인과 강도** — 78수 직후 AlphaGo의 큰 실수와 68수가 그 오류를 직접 만들었다는 해석을 분리했다.
3. **빈삼각의 위치** — 제2국의 공식 승착인 흑160 절단과, 신진서가 승리를 감지한 흑180 빈삼각을 구분했다.
4. **제3국 전략** — 제2국의 복잡한 중앙 전투와 제3국의 수비 지향적 안정 운영을 분리했다.
5. **승률의 정의** — 화면의 승률을 엔진·모델·규칙·덤·탐색 설정에 의존하는 평가치로 한정했다.
6. **KataGo의 계보** — AlphaZero형 자기대국과 신경망 유도 탐색을 효율화한 시스템으로 기술하고, 원 논문의 약 50배 계산 효율 보고를 반영했다.

출처 원장

- **S01 · AlphaGo 원 논문** — David Silver 외, “Mastering the game of Go with deep neural networks and tree search,” *Nature* 529, 484–489, 2016. <https://www.nature.com/articles/nature16961>
- **S02 · KataGo 원 논문** — David J. Wu, “Accelerating Self-Play Learning in Go,” 2019. <https://arxiv.org/abs/1902.10565>
- **S03 · KataGo Analysis Engine 문서** — https://github.com/lightvector/KataGo/blob/master/docs/Analysis_Engine.md
- **S04 · KataGo GTP Extensions 문서** — https://github.com/lightvector/KataGo/blob/master/docs/GTP_Extensions.md
- **S05 · Google DeepMind AlphaGo 공식 기록** — <https://deepmind.google/research/alphago/>

- S06 · 이세돌 인터뷰 — 중앙일보, “이세돌의 고백 ‘신의 한 수, 78수 아닌 68수,’” 2025-08-22. <https://v.daum.net/v/20250822050153568>
- S07 · 이세돌 강연 보도 — 연합뉴스, “이세돌 ‘알파고 꺾은 건 인간적 한 수 덕분,’” 2026-04-09. <https://www.yna.co.kr/view/AKR20260409121600017>
- S08 · 이세돌 강연 상세 보도 — 디지털조선일보, “AI 시대, 한 번도 내려본 적 없는 결정 내려야,” 2026-06-24. https://digitalchosun.dizzo.com/site/data/html_dir/2026/06/24/2026062480127.html
- S09 · 대회 규정 — 한국기원, “신진서 vs 카타고, 인간과 AI의 대결 앞두고 미디어데이 열려,” 2026-07-16. https://www.baduk.or.kr/news/report_view.asp?news_no=5870
- S10 · 제1국 공식 기록 — 한국기원, “신진서, 바둑 AI 카타고에 1국 패배,” 2026-07-17. https://www.baduk.or.kr/news/report_view.asp?news_no=5872
- S11 · 제2국 기록과 인터뷰 — 연합뉴스, “신진서 ‘중양 절단 수가 승착,’” 2026-07-19. <https://www.yna.co.kr/view/AKR20260719040200007>
- S12 · 최종 결과 — 한국기원, “인간대표 신진서, 카타고에 최종 승리,” 2026-07-21. https://www.baduk.or.kr/news/report_view.asp?news_no=5878
- S13 · 제2국 빈삼각 — 한국경제, “전투는 자살행위’라던 신진서, 결국 AI와 난전에서 웃었다,” 2026-07-19. <https://www.hankyung.com/article/202607192931i>
- S14 · 제3국 기자회견 — 연합뉴스, “신진서 ‘인간 바둑은 AI와 달리 승부수와 묘수 던지는 게 매력,’” 2026-07-21. <https://www.yna.co.kr/view/AKR20260721149400007>
- S15 · 프로젝트 최초 인터뷰 포인터 — 지식인사이드, “중국이 한국 바둑 1위를 넘지 못하는 이유 — 신진서 X 이세돌,” 2026-08-07. <https://www.youtube.com/watch?v=BZ1vYl27P9E>

판본 계보

- 부모 마스터: GO_AI_LIFE_COSMOS_05_PROSE_MASTER_v0.1
- 검증 보고서: GO_AI_LIFE_COSMOS_06_FACT_VERIFICATION_PASS01_v0.1
- 사실검증 통합본: GO_AI_LIFE_COSMOS_05_PROSE_MASTER_v0.2
- 현재 출판 정본: GO_AI_LIFE_COSMOS_07_PUBLICATION_CANONICAL_MASTER_v1.0

편집 정본 잠금과 전체 사실검증 완료는 다르다. 검증 범위 밖의 철학·인지과학·사회철학·문명론 문헌은 후속 PASS에서 별도로 검증하며, 외부 배포물은 이 제한을 이어받는다.

FILE LOCK

문서 ID	GO_AI_LIFE_COSMOS_07_PUBLICATION_CANONICAL_MASTER
버전	v1.0
부모 마스터	GO_AI_LIFE_COSMOS_05_PROSE_MASTER_v0.2
통합 보고서	GO_AI_LIFE_COSMOS_06_FACT_VERIFICATION_PASS01_v0.1
SOURCE FREEZE	2026-09-23
생성일	2026-09-24
PDF 분량	182쪽
검증 상태	PASS 01 INTEGRATED / REMAINING DOMAINS PENDING
외부 배포	정본 직접 배포 HOLD / 승인된 파생본만 사용
보존 식별	PDF SHA-256은 동반 METADATA에 기록

이 판본은 바둑·AI 사례의 사실, 당사자의 회고, 프로젝트의 해석을 분리한다. 후속 파생물은 이 구분과 HARD LOCK을 유지하며 부모 판본을 명시한다.