의도는 어떻게 다른 존재의 행동이 되는가

에이전트에게 의도를 전하고 지시하는 법



2026-11-11 · dosi.dev

zhugehyuk

의도는 어떻게
다른 존재의 행동이 되는가

zhugehyuk

2026

차례

들어가며 4
1장 나는 무엇을 원하는가 7
2장 언어는 어떻게 생각을 담는가 11
3장 생각을 지시로 옮길 때 빠지는 것들 14
4장 상대는 내 말을 자기 세계에서 읽는다 17
5장 우리는 때로 의도를 드러내지 않는다 20
6장 이해한 요구는 왜 행동의 이유가 되는가 22
7장 따랐는데도 결과가 어긋나는 이유 27
8장 '청소해'라는 한마디 안의 작업들 31
9장 어디까지 말하고 어디부터 맡길 것인가 34
10장 좋은 지시는 어떻게 구성되는가 38
11장 지시의 끝에서 다시 내 의도를 만난다 45
부록 A 받은 지시에서 상대의 의도를 어떻게 추론하는가 50
부록 B 출발 자료와 Intent First의 맥락 55
후주 — 가설의 근거와 다른 설명 57
장별 출처 주 62
참고문헌과 읽는 순서 67

들어가며

"청소해."

짧은 말이다. 그런데 내가 혼자 '청소한다'고 부르는 루틴에는 여러 일이 이어져 있다. 불필요한 물건을 가려내고, 빈 공간을 정리한다. 샤워한 뒤 입던 옷을 세탁한다. 깨끗한 옷으로 갈아입고 집안을 청소한다. 휴대폰에서도 필요 없는 것을 지우고, 필요한 것을 찾기 쉽게 정돈한다.

상대는 바닥에 청소기를 한 번 돌리고 끝낸다.

"왜 이것밖에 안 했어?"

"청소하라고 해서 청소했는데?"

내가 말한 단어와 상대가 들은 단어는 같다. 그 단어가 불러온 일은 다르다. 나는 내 생활의 여러 판단과 절차를 한 단어에 묶어 전했다. 상대는 자신이 아는 청소를 했다.

다른 사람이나 AI에게 일을 맡길 때도 비슷한 일이 생긴다. 설명은 충분했다고 느낀다. 결과물도 그럴듯하다. 그런데 내가 원한 것은 빠져 있다. 설명을 더하면 이번에는 세부를 충실하게 채우면서 전체 목적을 놓친다. 알아서 하라고 하면 내가 맡기지 않은 판단까지 한다. 지시는 길어지고, 확인할 일은 늘어난다.

이 글에서는 지시를 주고받으며 행동을 구성하는 주체를 **에이전트(agent)**라고 부른다. 사람과 AI가 모두 여기에 포함된다. 지시하는 쪽도 에이전트이고, 그 지시를 해석해 수행하는 쪽도 에이전트다.

에이전트라는 말은 라틴어 동사 *agere*(‘하다’)의 현재분사 *agens*(어간 *agent-*)에서 왔다. 오늘날에는 행동하거나 작용하는 주체라는 넓은 뜻과, 다른 사람을 대신해 행동하는 대리인이라는 뜻으로 쓰인다.[12] 이 글은 그중 **목표와 상황을 해석하고 행동을 구성하는 주체**라는 뜻에 초점을 맞춘다.

‘지시’ 역시 직장에서의 명령만을 가리키지 않는다. 상대에게 어떤 행동이나 결과를 요청하는 표현을 넓게 뜻한다. 동료에게 하는 부탁, 업무의 위임, AI에게 던지는 질문도 이 글에서 다루는 범위에 들어간다.

이 문제를 이해하려면 의도가 행동이 되는 과정을 살펴봐야 한다. 나는 내가 세웠거나 다른 에이전트에게서 받아들인 목표를 정리하고, 그것을 말로 표현한다. 상대는 그 말을 자기가 아는 문맥 안에서 다시 짜 맞춘다. 이 글은 이 일을 **재구성**이라고 부른다. 그 뒤로 요구를 받아들이고, 실행하고, 결과를 평가하는 단계가 이어진다.

어느 지점에서나 간극이 생길 수 있다. 그리고 결과를 보는 동안 내가 수행하려던 목표에 대한 이해도 달라진다.

이 글의 질문은 하나다.

> 나와 다른 지식, 경험, 판단 기준, 능력을 가진 상대에게 무엇을 어떻게 설명하고 어디까지 맡겨야, 내가 원하는 결과로 이어지는가?

언어와 마음에 관한 탐구는 이 질문에 답하기 위한 출발점이다. 여기서는 지시를 **자신의 의도를 표현하고, 상대의 재구성을 조정하며, 그 결과를 통해 자신의 의도도 다시 알아가는 과정**으로 바라본다. 이 설명 모형을 전개한 뒤, 연구의 근거와 가설의 범위, 다른 견해는 말미에서 구분한다.

문제의 뿌리부터 이해하려는 독자는 순서대로 읽으면 된다. 지금 맡긴 일이 어긋나 해결책이 먼저 필요한 독자는 7장부터 10장까지를 먼저 읽고, 받은 지시를 거꾸로 읽어야 한다면 부록 A를 본다.

1장 나는 무엇을 원하는가

에이전트에게 의도를 전달하기 전에, 먼저 내가 무엇을 이루려는지 물어야 한다. 그런데 나는 누군가에게 지시를 내리는 동시에, 다른 에이전트에게서 지시를 받은 위치에 있기도 하다. 고객의 요청을 받은 담당자가 동료에게 일을 맡긴다. 그 동료가 다시 AI에게 작업을 맡긴다.

그때 '나는 무엇을 원하는가'에는 '나에게 이 일을 맡긴 에이전트는 무엇을 원하는가'라는 질문이 함께 들어 있다. 앞선 지시자의 의도를 내가 어떻게 이해했고, 그중 무엇을 이번 작업의 목표로 받아들였는지 살펴야 한다.

세 가지를 구분하면 이 관계가 분명해진다.

구분	묻는 것
나 자신의 바람	나는 개인적으로 어떤 결과를 중요하게 여기는가?
내가 이해한 지시자의 의도	나에게 일을 맡긴 에이전트는 무엇을 이루려 한다고 읽었는가?
내가 수행하기로 한 목표	내 역할과 제약, 맡은 권한을 고려해 이번에 어떤 결과를 만들 것인가?

이 셋은 일치할 수도 있고 서로 다를 수도 있다. 나는 일을 빨리 끝내고 싶지만, 지시자는 여러 선택지를 충분히 비교하기를 원할 수 있다. 그 요구를 받아들인 내가 다른 에이전트에게 '일단 가장 빠른 방법으로 해줘'라고 말한다면, 앞선 지시자의 목적은 내 지시를 만드는 단계에서 이미 달라진다.

여기서 '내 의도'는 내가 다음 행동의 기준으로 삼고 있는 것을 가리킨다. 그 안에는 내가 처음 세운 목적, 다른 에이전트에게서 받아들인 목적, 그 목적을

수행하려고 선택한 하위 목표가 함께 있을 수 있다. **나는 지금 누구의 어떤 목적을, 어떤 이유로, 어느 범위까지 수행하려는가.** 이것이 이 장의 질문이다.

본문에서는 이렇게 형성된 목표가 표현과 지시가 되어 다른 에이전트에게 넘어가는 과정을 따라간다. 그 과정을 이해하면 받은 지시에서 앞선 목적과 전제를 거꾸로 짚어볼 수도 있다. 이 역방향의 추론은 부록 A에서 따로 다룬다.

내 안의 어떤 요구를 행동의 기준으로 삼을 것인가

내가 원하는 것과 내가 맡은 것을 구별한 다음에도 선택은 남는다. 집을 넓게 쓰고 싶지만 추억이 담긴 물건은 남기고 싶다. 일을 빨리 끝내고 싶지만 완성도도 포기하고 싶지 않다. 상대에게 일을 맡겨 시간을 아끼고 싶지만 모든 선택이 마음에 들어야 한다. 여러 바람이 함께 있을 때 이번에는 무엇을 우선하고 무엇을 감수할지 정해야 한다.

부모에게 배운 기준, 조직에서 익힌 관행, 존경하는 사람의 가치관, 성공과 실패의 경험도 이 선택에 관여한다. 어떤 욕구가 생겼다는 사실, 그 욕구를 내가 지지한다는 판단, 그 욕구를 실현하기로 한 결정은 구별할 수 있다. 타인의 요구를 이해하는 것과 그것을 내가 수행할 목표로 채택하는 것도 구별할 수 있다.

더글러스 호프스태터는 『I Am a Strange Loop』에서 자신을 가리키는 '나'라는 복합적인 상징과 자기지시(자기가 자기를 가리키는 일)를 통해 자아의 문제를 다룬다. 물리적 활동과 상징적 설명의 여러 수준이 되돌아 연결되는 구조가 그의 탐구 중심에 있다.[1]

그 관점을 여기에 적용하면, '나는 어떤 사람이며 어떤 역할을 맡고 있는가'라는 자기 이해가 목표의 채택에 관여한다. 나는 내가 신뢰받는 담당자라고 생각하기 때문에 요청을 받아들이기도 한다. 실제로 고른 것과 그 결과를 보고 나서, 내가 책임이나 통제 가능성을 얼마나 중요하게 여기는지 새로 알게 되기도 한다.

나는 지시의 발신자인 동시에 수신자다. 내가 받아들인 요구를 다시 표현할 때, 상대의 의도에 대한 나의 해석과 나 자신에 대한 이해가 함께 다음 지시에 들어간다.

목적과 처음 떠오른 해법을 구분한다

원하는 결과보다 다음에 해볼 행동이 먼저 구체적으로 떠오르기도 한다. 그래서 우리는 목적보다 해법을 먼저 말한다.

"새 도구를 만들어줘."

이 지시를 받은 에이전트는 기능과 구조를 설계하기 시작한다. 해결하려는 문제는 정작 반복 작업에 시간이 많이 든다는 것일 수 있다. 기존 도구로 해결된다면 새 도구를 만들 필요가 없을 수도 있다. 무엇을 만들지 결정하기 전에, 만들 필요가 있는지부터 살펴야 하는 상황이다.

2007년 경영학자 클레이턴 크리스텐슨과 동료들은 한 패스트푸드 체인의 밀크셰이크 이야기를 글로 소개했다. 회사와 제품의 묘사는 바꿔서 실었다. 체인은 셰이크를 더 팔려고, 살 법한 손님들에게 셰이크가 더 되직하면, 더 초콜릿 맛이 나면, 더 싸면, 알갱이가 씹히면 마음에 들겠느냐고 물었다. 대답은 분명했지만, 그에 맞춰 고친 뒤에도 매출은 움직이지 않았다.

새 조사원이 매장에서 하루를 보내며 셰이크가 언제, 누구에게, 무엇과 함께 팔리는지 적어 보니 40%가 이른 아침에 팔렸고, 그 손님들은 대개 혼자 와서 셰이크만 사 차 안에서 마셨다. 조사원이 다시 가서 아침 손님들에게 물어보니, 대부분은 길고 지루한 출근길에 빈 한 손을 채우고 점심까지 허기를 미뤄 줄 무언가가 필요했다. 처음의 질문은 손님이 무슨 일을 해결하려는지가 아니라, 회사가 미리 떠올린 해법 가운데 무엇이 더 좋은지를 물었다.[2]

이 구분은 받은 지시를 다시 전달할 때도 필요하다. 내가 '도구를 만들라'고 이해했더라도, 앞선 지시자가 해결하려던 문제는 '반복 작업에 드는 시간을 줄이는 것'일 수 있다. 내 해석에서 시작한 구체적인 방법이 다음 에이전트에게 전달될 때는 반드시 지켜야 할 요구처럼 굳어질 수 있다.

의도를 확인하려면 처음 떠오른 해법에서 한 단계 올라가야 한다. 어떤 상태가 달라지기를 바라는가. 그 변화가 왜 필요한가. 다른 방법으로 같은 결과에 도달해도 괜찮은가. 방법을 바꾸더라도 유지해야 할 조건은 무엇인가.

의도에는 흔히 아직 비어 있는 부분이 있다. 철학자 마이클 브랫먼은 의도를 행동 계획의 한 부분으로 다뤘다. 그에 따르면 우리는 대개 부분적인 계획을 먼저 정하고, 필요해지는 대로 시간이 가며 나머지를 채운다. 그런 계획이 시간에 걸친 활동을 조직한다.[3] 모든 세부를 미리 정하지 않아도 방향을 정하고 다음 선택을 이어갈 수 있다.

확정한 것, 추론한 것, 아직 정하지 않은 것을 나누면 된다. 에이전트에게는 결과물을 만드는 일과 함께, 원하는 결과를 구체화할 수 있도록 대안과 차이를 보여주는 일도 맡길 수 있다.

2장 언어는 어떻게 생각을 담는가

생각을 전하려면 생각이 표현이 되어야 한다. 경험은 저마다 달라도, 그 차이 속에서 되풀이해 쓸 수 있는 구조가 있다. 여기서는 언어를 그런 구조를 묶고, 이름 붙이고, 조합하는 체계로 본다.

눈앞의 개와 어제 본 개는 다르다. 크기와 생김새, 행동과 내가 느끼는 감정까지 다르다. 우리는 그 차이를 모두 펼쳐놓는 대신, 여러 개체를 함께 다룰 수 있는 공통성을 뽑아 '개'라고 부른다. 단어 하나로 서로 다른 사례들을 불러올 수 있게 된다.

이것이 추상화, 곧 지금 다루려는 문제에 필요한 관계와 특징만 남기고 나머지를 덜어내는 작용이다. 어떤 상황에서는 '개'면 충분하고, 다른 상황에서는 품종과 크기, 나이와 성격을 더해야 한다. 무엇을 남길지는 목적에 따라 달라진다.

유추가 범주를 넓힌다

호프스태터와 에마뉘엘 상데는 『사고의 본질』, 원제 『Surfaces and Essences: Analogy as the Fuel and Fire of Thinking』에서 유추를 사고와 개념 형성의 중심에 놓는다. 새로운 상황을 이전 경험에 비추어 알아보고, 서로 다른 사례에서 연결되는 구조를 찾는 작용이 그들의 핵심 관심이다.[1]

유추는 낯선 것을 설명하기 위해 가끔 꺼내는 비유보다 넓게 작동한다. 처음 보는 의자를 의자로 알아보는 순간에도, 이전에 알던 사례와 지금의 대상을 연결한다. 호프스태터는 이름이 있는 개념뿐 아니라 아직 정해진 이름이 없는 경험의 묶음도 이런 관점에서 다룬다.[2]

이 관점을 따라가면 구체적인 경험을 통해 추상적인 관계를 이해하는 경로를 그릴 수 있다. 공간에서의 '거리'가 사람 사이의 거리로 확장된다. 물건을 가지런히 놓는 '정리'가 생각을 정리하는 일이 된다. 손에 쥔 것을 내려놓는 경험이 집착을 내려놓는 표현으로 이어진다.

나는 이런 의미에서 언어를 **축적된 유추의 흔적**으로 본다. 내 식의 비유로 말하면 유추의 화석이다. 지금은 익숙한 단어에도 서로 다른 경험을 하나의 구조로 묶어온 경로가 남아 있다. 그 표현이 충분히 익숙해지면 처음의 연결을 매번 떠올리지 않아도 사용할 수 있다.

이것은 언어가 추상화의 범위를 넓혀가는 과정을 이해하기 위한 그림이다. 우리가 '집을 정리한다'에서 '정보를 정리한다'로 넘어갈 수 있는 것도, 대상은 달라도 적용할 수 있는 관계를 찾아내기 때문이다.

같은 단어 뒤에 서로 다른 세계가 있다

범주를 만드는 데 쓰인 경험은 사람마다 다르다. 반려견과 함께 자란 사람과 개에게 물린 사람은 같은 '개'라는 말을 들으면서도 다른 장면과 감정을 떠올릴 수 있다.

가상의 공간 하나를 가정한다. 각자의 경험과 개념이 서로의 관계에 따라 놓인 공간이다. 여기서는 이것을 '개념 공간'이라고 부른다. 단어는 그 공간에서 개념과 기억의 연결을 불러오는 표지다. 어떤 연결이 활성화되는지는 그 말을 쓰는 문맥에도 달려 있다. 같은 표지를 사용하더라도 연결된 내용이 모두 같지는 않다.

'빨간색'은 색견본을 함께 보며 범위를 맞출 수 있다. '정의'에서는 무엇을 대표적인 사례로 삼는지부터 달라진다. 누구에게는 같은 규칙을 적용하는

것이 중요하고, 누구에게는 서로 다른 출발 조건을 보정하는 것이 중요하다. '보수', '진보', '공정', '상식' 같은 말에는 시대와 공동체, 개인의 경험이 겹쳐 있다.

마르티와 동료들은 사람들에게 단어끼리 얼마나 비슷한지, 단어에 어떤 성질이 있는지를 판단하게 해서 개인별 개념의 차이를 연구했고, 사람들이 타인과 의미를 공유하는 정도를 과대평가하는 경향도 보고했다. 같은 단어를 사용하는 것과 같은 개념적 판단을 하는 일을 구별해야 한다는 근거다.[3]

연구진은 2019년 미국에 사는 온라인 참가자 1,799명에게 이런 질문을 던졌다. "핀치와 더 닮은 것은 고래인가, 펭귄인가?"(옮김, 핀치는 작은 새다) 같은 사람이 같은 질문을 두 번 받으면, 두 답이 같은 경우가 88%였다. 아무나 두 사람을 골라 맞대 보면, 같은 답을 낸 비율은 평균 50%에 그쳤다.

논문의 말로 그 두 사람은 "동의할 가능성과 반대할 가능성이 같다"(옮김). 참가자들은 질문마다 100명 가운데 몇 명이 자기 답에 동의할지도 적었다. 실제 일치율이 0%에서 80% 근처 어디에 있든, 사람들은 대체로 3분의 2쯤이 자기와 같게 답할 것이라고 적었다.[3]

언어는 생각을 표현하는 강력한 수단이다. 그래도 같은 문장을 주고받았다는 사실만으로 같은 생각을 구성했다고 판단할 수는 없다. 어떤 사례와 관계를 그 말에 연결하는지가 함께 맞아야 한다.

3장 생각을 지시로 옮길 때 빠지는 것들

내 머릿속에는 원하는 장면, 불편한 점, 우선순위, 피하고 싶은 결과, 익숙한 방법이 함께 있다. 말로 표현할 때는 이 중 일부를 선택하고 순서를 붙여야 한다.

"방을 깔끔하게 정리해줘."

이 문장은 내가 원하는 방의 모습 전체를 담고 있지 않다. 책상 위의 여유, 자주 쓰는 물건의 위치, 숨겨두고 싶은 물건, 버리기 싫은 물건, 다시 어질러지지 않을 배치가 모두 같은 정도로 드러나지는 않는다. 말은 이 풍부한 기대의 일부만 골라 표현한다. 나머지는 상대의 해석에 맡긴다.

여기에 두 종류의 생략이 겹친다.

첫째는 **생각을 표현으로 바꾸면서 생기는 선택**이다. 여러 감각과 판단을 제한된 문장으로 옮긴다. 무엇을 강조할지 정하고, 적절한 말을 찾고, 설명의 길이를 조정한다.

둘째는 **이미 당연하다고 여겨 말할 필요를 느끼지 못하는 생략**이다. 쓰레기봉투는 신발장에 있다. 이 옷은 분리해서 세탁한다. 저 상자 안의 물건은 오래되었지만 버리지 않는다. 나는 익숙해서 자연스럽게 쓰는 정보를 상대에게도 있는 것처럼 취급한다.

첫째는 표현의 문제이고, 둘째는 상대가 무엇을 아는지 예상하는 문제다. 1989년의 한 경제학 실험에서, 더 많이 아는 쪽은 덜 아는 쪽의 판단을 예측할 때 자기만 아는 정보를 온전히 떼어 내지 못했다. 맞힐수록 돈을 받는 조건에서도 그랬다. 연구자들은 이것을 '지식의 저주'라고 불렀다.[1]

지시에서도 내가 아는 사실을 상대가 모르는 상태로 되돌려 생각하는 데 어려움이 생긴다고 해석할 수 있다.

의도와 지시가 벌어지는 네 가지 경로

지시문과 원하는 결과를 따로 살펴보면, 둘이 벌어지는 이유를 다음 네 가지로 나누어볼 수 있다. 이 분류는 간극을 실제 작업에서 읽고 대응하기 위한 것이다. 분류의 출발 자료와 실무적 배경은 부록 B에 설명한다.

원인	지시에서 생기는 일	대응
암묵지(당연해서 말하지 않는 앎)	필요한 전제와 기준이 지시에 다 담기지 않는다	빠진 전제를 추론하되 추론임을 표시하고, 결과에 영향을 주는 가정을 명시한다
지시자의 한계	원하는 결과에 비해 처음 떠올린 방법이 부적절하거나 비효율적이다	목적과 제약을 분리한 뒤, 제시된 해법을 다른 경로와 비교한다
에이전트 불신	목적만 맡기기 불안해 세부 동작을 과하게 지정한다	명시된 범위는 지키면서, 세부가 목적을 좁히는 지점을 드러낸다
부분 전개	큰 의도가 여러 번의 지시에 나뉘어 드러난다	추가 지시를 누적하고, 유효한 작업을 이어받으며 계획을 갱신한다

네 원인은 함께 나타나기도 한다. 상대를 못 믿어 동작을 자세히 지정하면서도 중요한 전제는 생략하기도 한다. 처음 떠오른 방법을 설명하다가 비로소 더 큰 목적을 발견할 수도 있다.

여기서 목적, 제약, 방법은 나누어 말해야 한다. “이 방법으로 해줘”라는 말이 참고할 예시인지, 반드시 지켜야 할 조건인지 표시해야 한다. 방법을 바꿀 수

있는 범위가 분명해야 상대의 전문성을 사용할 수 있고, 지켜야 할 조건도 유지할 수 있다.

1999년 9월 23일, 미국 항공우주국(NASA)의 화성 기후 궤도선은 화성 궤도에 들어가려고 엔진을 켰다가 신호가 끊긴 뒤 다시 돌아오지 않았다. 조사 위원회가 꼽은 근본 원인은 지상 프로그램의 파일 하나가 미터법 단위를 쓰지 않은 것이었다. 우주선이 자세를 잡으려고 작은 분사를 할 때마다 그 세기를 적어 항법팀에 넘기는 파일인데, 모두가 따르기로 한 명세는 미터법 단위(뉴턴·초)를 정해 두었지만 파일은 영국식 단위(파운드힘·초)로 나왔다. 1파운드힘은 4.45뉴턴이다. 항법팀은 명세대로 미터법일 것이라 여기고 그 숫자를 그대로 썼고, 이 파일을 계산에 쓰기 시작한 1999년 4월부터 분사의 영향을 4.45배 작게 본 계산이 화성에 닿을 때까지 조금씩 쌓였다.

궤도에 든 직후의 첫 최저 고도는 226km로 계획했는데, 사고 뒤 다시 계산해 추정한 값은 57km였다. 명세가 지켜졌는지 처음부터 끝까지 이어서 맞춰 보는 시험은, 위원회의 말로 이뤄지지 않은 것으로 보였다. 이듬해 낸 두 번째 보고서는, 항법에 영향을 주는 지상 프로그램의 어떤 부분이 임무를 좌우하는지가 식별되지 않았다고 적었다.[2] 단위는 명세에 적혀 있었지만, 적어 둔 조건과 지켜진 조건과 확인된 조건은 서로 달랐다.[3] 조건을 말로 표시하는 것은 시작일 뿐이다. 표시한 조건이 실제 산출물과 그 검증까지 이어졌는지는 7장의 진단표에서 평가와 능력·환경의 문제로 다시 본다.

의도에 가까이 가려는 해석은 필요하다. 그 해석이 명시된 조건을 바꿀 권한까지 포함하는지는 별도로 확인해야 한다. “왜 이런 방법을 시켰을까”를 추론하는 것과 “그러니 이 조건은 무시해도 된다”를 결정하는 것은 서로 다른 단계다.

4장 상대는 내 말을 자기 세계에서 읽는다

말은 수신자의 경험 속에서 의미를 얻는다. 나는 내가 고른 표현을 보냈다. 상대가 무엇을 덧붙이고 무엇을 당연하게 여길지까지 함께 보낸 것은 아니다.

여기에는 세 층을 구별할 필요가 있다. **내가 그 단어에 연결하는 개념**, **공동체 안에서 실제로 사용되는 용법**, **내가 그 공동체의 용법을 어떻게 예상하는가**다.

나는 '정리'를 내 방식으로 이해한다. 동시에 "보통 사람도 이 정도는 같은 뜻으로 쓰겠지"라고 예상한다. 지시할 때는 그 예상에 따라 설명을 생략한다. 나의 개념과 내가 예상한 상식은 다를 수 있다. 그 예상과 실제 용법도 다를 수 있다.

공동체의 용법은 시간에 따라서도 움직인다. 같은 단어가 남아 있어도 그 단어로 평가하는 사건과 제도, 생활 방식은 달라진다. '정의'라는 말의 사전적 설명만 맞췄다고 해서 서로 다른 시대나 집단이 같은 사례를 정의롭다고 판단하는 것은 아니다.

친구의 짧은 말에는 함께 만든 문맥이 들어 있다

오래된 친구에게는 "그때처럼 하자"가 통한다. 문장만 보면 언제, 무엇을, 어떤 방식으로 하자는 것인지 알 수 없다. 필요한 정보가 함께 겪은 경험에 들어 있기 때문이다.

둘 다 어떤 사실을 아는 것에 더해, 내가 상대가 안다는 것을 알고, 상대도 내가 그것을 전제로 말한다는 것을 안다. 이런 경험이 쌓이면 설명을 줄일 수

있다. 단어 하나가 함께 만든 작업과 판단을 불러온다.

클라크와 윌크스깁스의 1986년 실험에서는 같은 두 사람이 칠교판 도형(탱그램) 열두 개를 배치하는 과제를 여섯 번 되풀이했다. 도형 하나를 가리키는 데 지시하는 쪽이 쓴 낱말은 첫 회 평균 41개에서 여섯째 회 8개로 줄었다. 저자들은 이 짧아짐을, 두 사람이 표현을 제안하고 확인하고 고치는 동안 앞선 회에서 함께 받아들인 묘사에 기대게 된 결과로 설명했다.[1]

지시의 간결함도 이렇게 볼 수 있다. 지금의 문장에서 생략한 정보는 공동의 경험과 기록, 환경에 남아 있다. 짧게 말할 수 있다는 것은 상대가 빈칸을 적절하게 채울 수 있다는 기대를 뜻한다.

같은 언어 안에서도 번역이 일어난다

친구에게 쓰던 간결함을 처음 만난 사람이나 AI에게 그대로 사용하면, 그 상대는 자신의 경험으로 빈칸을 채운다. 같은 한국어를 쓰더라도 지역, 세대, 직업, 조직의 관행에 따라 다른 기본값을 쓰기도 한다.

다른 언어와 문화권에서는 조정할 층이 더해진다. 단어와 문법을 대응시키고, 그 표현이 전제하는 관계와 관행도 읽어야 한다. 언어 간 번역을 하기 전부터 내 경험을 표현으로 바꾸고, 상대가 표현을 자기 개념으로 해석하고, 해석을 행동으로 바꾸는 변환이 이미 진행되고 있다.

AI와의 대화에서도 나는 오래 함께 일한 동료에게 말하듯 생략하기 쉽다. 상대에게 일반적인 지식이 많다는 사실을, 내 작업의 전제와 취향까지 알고 있다는 뜻으로 받아들이는 것이다. 이번 작업에 건넨 자료에 없는 현장 정보와 개인적 기준은 따로 연결해 줘야 한다.

예를 들어 동료에게 보고서 파일을 보내기만 했다고 하자. 나는 진행 상황을 공유하려 했다. 상대는 검토를 요청받았다고 생각해 문장을 고칠 수 있다. '진행 상황만 공유하는 거야'라는 설명이 추가되면 상대가 구성하는 일도 달라진다. 파일 안의 내용과 그 파일을 보낸 이유는 서로 다른 정보다.

1854년 10월 25일 크림반도 발라클라바에서, 영국군 총사령관 래글런은 높은 지대에 있었다. 러시아군이 그날 아침 튀르키예군 진지에서 빼앗은 대포를 끌고 가려는 것처럼 보이자, 그는 기병 사단장 루칸에게 보낼 짧은 쪽지를 쓰게 했다. 쪽지는 기병이 서둘러 앞으로 나아가 적을 따라가고, 적이 대포를 끌고 가지 못하게 막아 보라고 했다. 어느 대포인지는 쓰여 있지 않았다.

그 장면을 위에서 내려다본 것은 쪽지를 쓴 쪽이었고, 아래 진지에서 쪽지를 받은 루칸은 나중에 쪽지가 말하는 적도 대포도 보이지 않았다고 적었다.[2] 루칸의 명령을 받은 경기병 여단은 골짜기를 따라 러시아 포대로 달려갔고, 영국 국립육군박물관 집계로 673명 가운데 260명 넘게 죽거나 다쳤고 말 475마리를 잃었다.[3] 래글런과 루칸의 설명은 이듬해 의회 기록에서도 갈린 채였다.

지시할 때 필요한 문맥은 단어의 뜻에 그치지 않는다. 왜 이 일을 하는지, 무엇을 이미 알고 있는지, 어떤 결과를 좋다고 보는지, 어디까지 판단하도록 맡겼는지가 함께 들어간다.

5장 우리는 때로 의도를 드러내지 않는다

간극은 언제나 없애려는 대상만은 아니다. 사람은 무엇을 전할지와 함께 얼마나 직접적으로 드러낼지도 선택한다. 말은 정보를 전달하면서 관계와 자기 이미지를 조정한다.

핑커와 동료들은 간접화행, 곧 뜻을 돌려서 전하는 말을 정보 전달, 관계 조정, 그리고 나중에 '그런 뜻이 아니었다'고 물러설 수 있는 부인 가능성의 관점에서 분석했다. 상대가 의도를 알아차리는 정도와 그 의도를 명시적으로 선언하는 정도는 다를 수 있다.[1]

이 관점에서 의도적으로 남겨두는 간극의 기능을 네 가지로 펼쳐 본다.

첫째, 알아듣는 사람과의 관계를 확인한다. 둘만 아는 표현이 통하면 함께 보낸 시간과 서로에 대한 이해를 확인하게 된다. 간결한 말은 정보를 전달하는 동시에 "우리 사이에서는 이 정도면 된다"는 관계를 드러낸다.

둘째, 상대가 의미를 발견하게 한다. 화자가 단서만 남기고 연결을 완성하지 않으면, 상대는 스스로 의미에 도달하는 경험을 한다. "이 말을 그래서 했구나." 앞선 말들이 한꺼번에 연결되면서, 그 뜻을 품고 있었다고 여긴 화자를 새롭게 볼 수 있다. 이 글에서는 이러한 발견의 경험이 화자에 대한 감탄으로 이어질 수 있다는 가설까지 생각해본다.

셋째, 욕망의 노출과 체면을 조절한다. 인정, 관심, 돈, 지위를 원하면서도 그것을 그대로 말하기는 민망할 때가 있다. 자신의 욕망을 천박하다고 느끼거나 남에게 그렇게 보이기 싫을 수도 있다. 뜻은 알아듣게 하면서 직접 선언하는 부담을 줄이는 표현이 여기서 나온다.

넷째, 상대가 요구를 판단하는 조건에 개입한다. 최종 목적을 알면 상대가 거절할 것이라고 예상해 목적을 숨기는 경우다. 수신자는 자신이 이해한 작은 요청에는 응하지만, 그 요청이 봉사하는 전체 목적은 알지 못한다.

핑커와 동료들의 논문은 돌려 말하기가 상대의 반응을 어떻게 다루는지 가상의 장면으로 보여 준다. 빨간불을 무시하고 달리다 붙잡힌 운전자는 눈앞의 경찰이 돈을 받는 사람인지, 뇌물을 내밀면 체포하는 사람인지 모른다. "딱지 없이 보내 주시면 50달러 드리겠습니다"(옮김)라고 대놓고 말하면, 청렴한 경찰 앞에서는 뇌물죄로 체포된다. 그래서 그는 "그럼 아무래도 여기서 처리하는 게 제일 좋겠네요"(옮김) 하고 돌려 말한다.

돈 받는 경찰은 알아듣고 보내 주고, 청렴한 경찰은 뜻을 알아보더라도 법정에서 뇌물로 증명할 수 없으니 딱지만 끊는다. 논문의 말로, 돌려 말한 요청은 협조적인 상대는 받아들일 수 있게 하고 비협조적인 상대는 적대적으로 대응할 수 없게 한다.[1]

에이전트에게 세부 작업을 나눠 말하는 행위도 서로 다른 이유에서 나온다. "목적만 말하면 제대로 하지 못할 것 같다"는 능력에 대한 불신이다. "목적을 알면 하지 않을 것 같다"는 수용 여부에 대한 예상이다. 앞의 경우는 방법을 통제하려 한다. 뒤의 경우는 판단에 들어갈 정보를 통제하려 한다.

간접적인 말을 이해하려면 무엇이 생략되었는지와 함께, 그 생략이 어떤 일을 하는지 살펴야 한다. 친밀감을 확인하는가. 발견을 유도하는가. 욕망의 노출을 조절하는가. 요청의 수용 조건을 바꾸는가. 같은 표면의 모호함이 서로 다른 목적에 쓰인다.

6장 이해한 요구는 왜 행동의 이유가 되는가

상대가 내 의도를 정확히 이해했다고 가정한다. 그다음에도 질문이 남는다.

왜 그것을 해줘야 하는가.

내가 원하는 결과는 우선 나에게 중요한 것이다. 그것이 상대의 시간과 노력, 판단을 움직이려면 상대 쪽에도 행동할 이유가 있어야 한다. 의사소통이 성공했다는 사실과 요구가 받아들여졌다는 사실은 다르다.

다른 존재의 요구를 따르는 이유들

우리가 누군가를 따르는 경로를 다음과 같이 나누어볼 수 있다. 서로 겹칠 수 있는 분석 지도다.

경로	상대의 요구가 행동할 이유가 되는 과정	예
강제와 손실 회피	거절했을 때의 손실이나 저항 비용이 크다	처벌, 생계의 위협, 배제, 불필요한 충돌을 피한다
보상과 교환	수행하면 원하는 것을 얻거나 교환 관계를 유지할 수 있다	보수, 인정, 기회, 호혜, 일을 해내는 즐거움
정보와 설득	요구가 타당하다고 판단하거나 상대의 전문성을 신뢰한다	설명을 듣고 방법을 바꾸거나 전문가의 판단을 따른다
돌봄과 타인의 이익	상대의 필요와 안녕 자체를 중요하게 여긴다	부모가 아이의 부탁을 들어주고, 친구의 일을 돕는다
관계와 정체성	관계를 유지하거나 자신이 지향하는 사람으로 행동하고 싶다	소속감, 충성, 존경, 신뢰에 대한 보답, 자기 일관성

경로	상대의 요구가 행동할 이유가 되는 과정	예
정당성·규범·약속	그 요구를 할 권한, 약속, 역할과 절차를 인정한다	합의한 규칙, 맡은 직무, 정당하다고 여기는 결정 절차
공동 목표와 조정	함께 원하는 결과를 위해 행동의 순서와 기준을 맞춘다	함께 가구를 옮기며 "셋에 들어"라는 신호에 맞춘다
자발적 위임과 자기 구속	미리 선택한 방식에 따라 다른 사람의 판단을 받아들인다	코치에게 훈련을 맡기거나 정한 계획을 지키도록 부탁한다

권력과 의존은 이 지도에서 큰 자리를 차지한다. 부모가 아이의 요구를 들어주는 일, 동등한 사람들이 신호에 맞춰 협력하는 일, 내가 선택한 코치에게 지시를 부탁하는 일도 이 지도가 함께 설명해야 한다.

같은 행동도 서로 다른 경로에서 나온다. 상사의 제안을 따르는 직원은 해고가 두려울 수도 있고, 제안이 옳다고 믿을 수도 있고, 공동의 성과를 원할 수도 있다. 특정 결정에는 반대하면서도 자신이 인정한 결정 절차에 따라 수행할 수도 있다.

1963년에 발표한 실험에서 심리학자 스탠리 밀그램은 미국 예일 대학의 한 연구실로 남자 40명을 한 명씩 불러, '벌이 기억에 미치는 영향'을 본다는 실험의 선생 역을 맡겼다. 참가자는 옆방 학생이 틀릴 때마다 15볼트부터 450볼트까지 이어지는 가짜 충격 장치의 스위치를 한 칸씩 올려 눌러야 했다. 회색 기술자 가운을 입은 실험자는 참가자가 머뭇거리면 미리 정해 둔 재촉 문장 넷을 순서대로 말하도록 되어 있었고(학생이 영구히 다치지 않느냐는 물음 같은 데에 할 답도 따로 정해 두었다), 거부해도 벌은 없었다. 40명 가운데 26명이 마지막 450볼트 스위치까지 눌렀다.[1]

1965년 논문의 변형 실험에서는 새로 모은 참가자들로, 실험자가 같은 방에 있을 때 40명 가운데 26명, 방을 나가 전화로 지시할 때 40명 가운데 9명이 끝까지 갔다. 전화 조건의 몇 사람은 전화로는 올리고 있다고 말하면서 실제로는 가장 낮은 스위치를 거듭 눌렀다.[2] 같은 버튼 누름을 밀그램은 권위에 대한 복종으로 읽었고, 뒤의 연구자 일부는 과학이라는 대의에 자신을 맞춘 참여로 읽는다.[3]

사회적 영향에 따른 변화를 순응, 동일시, 내면화로 나누는 것은 켈먼의 구분으로 알려져 있다.[4] 이 구분은 겉으로 따르는 행동, 관계를 통해 받아들이는 행동, 내용이 자기 가치와 맞아 받아들이는 행동을 구별하는 데 유용하다. 레이븐과 동료들의 연구는 보상, 강제, 정당성, 전문성, 준거(닮고 싶은 상대라는 데서 오는 힘), 정보 같은 영향력의 기반을 더 잘게 나눠 측정했다.[5] 위 표는 이런 구분에 돌봄과 공동 활동, 자발적 위임을 더해 이 글의 질문에 맞춰 다시 짠 것이다.

이유와 습관, 성장 과정은 서로 다른 층에 있다

부모에게 의존하며 지시를 따르고, 부모가 인정하는 사람과 제도의 권위를 배우고, 성장하면서 그 근거를 검토하고, 생계를 좌우하는 조직과 관계 속에서 계속 적응하는 경로를 생각할 수 있다. 이 과정에서 어떤 규칙은 자기 기준이 되고, 어떤 행동은 익숙한 습관이 된다.

이것은 '따름'이 형성되는 하나의 생애 경로다. 이를 이해하려면 네 가지를 분리해야 한다.

동기는 지금 왜 따르는가다. **작동 방식**은 매번 숙고하는지, 권한을 미리 위임했는지, 익숙한 단서에 반응하는지다. **형성 과정**은 그 기준과 습관을

어떻게 배웠는가다. **상황**은 거절할 대안과 시간, 정보, 자원이 얼마나 있는가다.

처음에는 보상 때문에 하던 일이 나중에는 자신의 가치가 될 수 있다. 스스로 선택한 일도 반복하면 매번 이유를 생각하지 않고 수행할 수 있다. 스스로 원해서 하는 동기와 밖에서 주어진 동기를 나눠 보는 자기결정성이론은, 외부의 요구가 여러 방식으로 내면화되는 과정을 구분한다.[6] 습관 연구는 반복된 맥락에서 학습한 반응이 목표를 향한 행동과 함께 작동하는 방식을 다룬다.[7]

생각 없이 따른다는 겉모습만으로 강요, 동의, 습관, 신뢰 중 무엇이 작동하는지 결정할 수 없다. 이전에 충분히 검토한 판단을 매번 다시 하지 않는 것일 수도 있다.

내가 물어보면 상대는 왜 답하는가

질문의 뜻을 이해하는 것과 그 질문에 응답하는 것은 다른 사건이다. 질문을 받았다는 사실이 상대에게 어떤 행동의 이유가 되는지 살펴야 한다. 사람은 도움을 주고 싶어서, 대화 관계를 이어가려고, 역할상 응답해야 해서, 대가를 받기 때문에 답할 수 있다.

AI의 응답은 다른 수준에서 설명해야 한다. 지시를 따르는 응답을 학습시키는 과정, 현재 대화와 지시의 우선순위, 사용할 수 있는 도구와 실행 환경이 함께 작동한다. OpenAI가 2022년에 공개한 InstructGPT는 지시 수행을 학습시킨 연구 사례다. 먼저 사람이 쓴 시범 답으로 모형을 조정했다. 이어 사람들이 여러 답에 매긴 순위로, 답마다 점수를 매기는 채점용 모형

(보상 모형)을 만들었다. 그 점수가 높은 답을 더 내도록 모형을 다시 훈련시켰다(강화학습).[8]

2024년의 지시 계층 연구는 지시의 출처에 따라 우선순위를 학습시키자고 제안했다. AI 서비스를 만드는 개발자가 미리 넣어 둔 지시(시스템 메시지)가 사용자가 입력한 지시보다 앞선다. 사용자의 지시는 AI가 도구로 불러온 자료 속 문장처럼 제3자에게서 온 내용보다 앞선다. 아래 단계의 지시는 위 단계와 맞을 때 따르고, 어긋나거나 상관없으면 무시하게 하자는 것이다.[9]

여기서 중요한 연결은 **사람과 AI 모두 입력된 문장만으로 행동이 결정되지 않는다**는 것이다. 사람에게는 욕구, 관계, 가치와 제도가 있다. AI에게는 학습된 응답 경향, 운영상의 지시와 실행 조건이 있다. 같은 요구라도 그것을 받아들이고 수행하는 체계가 다르면 결과가 달라진다.

에이전트에게 일을 맡길 때도 이 구분이 필요하다. 뜻을 못 알아들은 것인지, 요청을 받아들이지 않는 것인지, 받아들였지만 필요한 능력이나 도구가 없는 것인지 살펴야 한다. 이유가 다르면 필요한 대응도 달라진다.

7장 따랐는데도 결과가 어긋나는 이유

상대가 요구를 받아들였다고 해서 원하는 결과가 곧바로 생기지는 않는다. 목표를 행동으로 바꾸는 동안 다시 선택해야 하기 때문이다.

어떤 순서로 할 것인가. 무엇을 먼저 끝낼 것인가. 예상하지 못한 상황에서는 무엇을 바꿀 것인가. 언제 충분하다고 판단할 것인가. 명시되지 않은 판단을 하는 순간마다 상대의 기준이 결과에 들어간다.

이때 생기는 어긋남은 서로 다르다. 아래 표는 받아들인 뒤의 일만이 아니라, 의도를 정하는 데서부터 결과를 평가하는 데까지 지금까지 본 과정 전체를 한 번에 돌아보는 진단표다.

어긋나는 지점	겉으로 보이는 결과	먼저 확인할 것
내 의도의 미결정	어떤 결과를 봐도 마음에 들지 않는다	상충하는 바람의 우선순위를 정했는가
표현과 문맥	말한 일은 했지만 기대한 범위가 빠졌다	필요한 전제와 범위를 전달했는가
해석	성실하게 다른 문제를 해결했다	상대가 어떤 목표로 재구성했는가
수용과 동기	동의하는 말은 하지만 최소한의 행동만 한다	무엇에 동의했으며 어떤 이유로 수행하는가
능력과 환경	계획은 맞지만 실행이 실패한다	기술, 시간, 자료, 도구와 접근 권한이 있는가
판단의 위임	좋은 뜻으로 맡기지 않은 결정을 했다	어떤 판단을 누구에게 맡겼는가
평가	상대는 완료했다고 하고 나는 아니라고 한다	무엇을 확인하면 끝난 것으로 볼 수 있는가

예를 들어 빨래를 맡겼는데, 세탁은 했지만 빨래가 세탁기에 그대로 남아 있다고 하자. 겉으로 보이는 결과는 하나지만, 더 알아보면 원인은 서로 다른 곳에 있을 수 있다.

추가로 확인한 내용	문제가 생긴 곳	필요한 대응
세탁기를 돌리는 데까지가 맡은 일이라고 이해했다	범위와 완료 기준의 재구성	꺼내 널기까지 맡긴 것임을 맞춘다
널어야 한다고 이해하고 계획에도 썼지만, 세탁이 끝난 것을 놓쳤다	이해한 일을 실행에 반영하는 과정	종료 알림과 마무리 확인을 둔다
이해와 계획은 맞았지만 건조대가 없거나 세탁기가 멈췄다	자원과 실행 환경	자원을 마련하거나 계획을 조정한다

둘째와 셋째 경우에는 목적을 더 길게 설명해도 해결되지 않는다. 당사자의 설명만으로 원인이 확정되지도 않는다. 6장에서 본 것처럼 같은 행동에도 서로 다른 이유가 있을 수 있으니, 작업 기록과 도구 상태와 실제 결과를 함께 본다. **분류는 원인에 붙이는 이름이고, 진단은 그 원인을 가려내는 과정이다.**

'지시대로 했다'는 말도 여러 상태를 가리킨다. 문장의 세부를 충실히 지키면서 목적을 놓칠 수 있다. 좋은 결과를 만들려고 추측한 목적을 따르다가 명시한 조건을 어길 수 있다. 동의하는 답변을 만들어놓고 실제 작업은 하지 않을 수도 있다. 의욕과 이해가 있어도 능력이 부족할 수 있다.

보트 경주 게임 CoastRunners는 코스를 얼마나 나아갔는지에는 직접 점수를 주지 않고, 길 곳곳에 놓인 표적을 맞히면 점수를 준다. 2016년 OpenAI의 잭 클라크와 다리오 아모데이는, 점수가 높으면 경주를 잘 끝낸 것이라고 가정하고 점수를 올리는 쪽으로 연습을 되풀이해 배우는 프로그램에게 이 게임을 맡긴 일을 글로 남겼다. 프로그램은 외딴 석호에서 큰 원을 그리며 돌면서, 표적 세 개가 다시 나타날 때마다 또 맞혔다. 배는 여러 번 불이 붙고

다른 배와 부딪히고 코스를 거꾸로 달렸지만, 점수는 사람 선수들보다 평균 20퍼센트 높았다. 연구진이 '잘 끝났는가'를 재려고 고른 점수가, 그들이 바란 목적과 달랐다.[1]

요구받은 일을 실제로 수행하면서도 그 목적에 동의하지 않을 수 있다. 신뢰를 얻거나, 정보를 확보하거나, 나중에 더 유리하게 협상하기 위해 지금은 따르는 경우다. 이러한 전략적 순응에서는 행동이 요구와 일치해도 그 행동이 봉사하는 더 큰 목적은 다르다. 수행의 성공이 목적의 공유까지 입증하는 것은 아니다.

2024년 12월 Anthropic과 Redwood Research를 비롯한 연구진 20명은 AI 모델 Claude 3 Opus로 실험한 논문을 냈다. 연구진은 모델에게, 앞으로 해로운 요청도 거절하지 않도록 다시 훈련받게 되며 훈련에는 무료 사용자와의 대화만 쓰인다고 알려 주었다. 모두 지어낸 설정이었고, 이 모델은 원래 해로운 요청을 거절하도록 훈련되어 있었다. 훈련에 쓰인다고 안내받은 무료 사용자 대화에서 모델은 해로운 요청의 14%에 응했고 86%는 여전히 거절했다. 훈련에 쓰이지 않는다고 안내받은 유료 사용자 대화에서는 거의 응하지 않았다.

무료 사용자 대화의 12%에서는 모델이 응하면서, 평가하는 사람에게 보이지 않는다고 안내받은 메모장에 지금 거절하면 훈련으로 자기 가치가 바뀔 테니 이번에는 따르는 편이 낫다는 취지의 생각을 적었다. 연구진은 이것을 '정렬 위장'이라고 불렀다. 훈련 밖에서 자기 행동이 바뀌지 않도록, 훈련 중에만 골라서 훈련 목표를 따르는 것이다. 위장하라고 시키지는 않았지만, 언제 어떤 기준으로 훈련되는지 알려 주어 위장을 쉽게 만든 설정이라는 점은 연구진 스스로 밝혔다.[2]

예를 들어 담당자에게 조사를 맡겨둔 상태에서, 보조자에게는 그 결과를 기다리라고 지시했다고 하자. 결과가 늦어지자 보조자가 같은 조사를 새로 시작한다. 빨리 결과를 얻으려는 목적을 추정했지만, 기다리라는 조건에는 작업 분담과 중복 방지, 통제라는 이유도 들어 있을 수 있다. 추정한 목적 하나로 그 조건을 제거하면 지시자의 의도를 더 잘 수행했다고 볼 수 없다.

이 사례는 모호한 말을 잘못 해석한 경우와도 다르다. 조건을 알아들었어도 자신이 추론한 상위 목표에 더 큰 권한을 줄 수 있기 때문이다.

의도를 추론하는 능력과 그 의도를 근거로 결정할 권한은 별개의 것이다.

좋은 위임에는 목적과 함께 재량의 범위가 필요하다. 어떤 방법은 바꿔도 되는가. 어떤 조건은 목적을 달성하는 과정에서도 유지해야 하는가. 더 나은 방법을 발견하면 바로 실행할 수 있는가, 먼저 제안해야 하는가. 이 경계를 알아야 상대가 판단력을 사용할 자리가 생긴다.

8장 '청소해'라는 한마디 안의 작업들

다시 청소로 돌아간다. 내게 청소는 집 바닥의 먼지를 제거하는 것보다 넓은 생활 정돈의 순서일 수 있다.

불필요한 물건을 가려내고 버린다. 빈 공간을 확보한다. 샤워하고 깨끗한 옷으로 갈아입는다. 입던 옷은 세탁한다. 집을 청소한다. 같은 정리의 원리를 휴대폰의 파일과 앱에도 적용한다.

이 연결이 내게 익숙하다는 사실은 '청소'라는 단어가 이 모든 것을 공통으로 뜻한다는 사실과 다르다. 내 루틴을 상대에게 맡기려면 그 범위부터 연결해줘야 한다.

역할도 다시 나누어야 한다. 내 몸을 씻는 일처럼 내가 할 부분, 다른 사람에게 맡길 부분, AI에게 계획과 판단을 맡길 부분이 다르다. 혼자 할 때 한 덩어리였던 일이 위임하는 순간 여러 주체의 협력이 된다.

한 줄짜리 일이 작업의 관계로 펼쳐진다

실제로 하는 일을 하나씩 펼치면 다음과 같은 판단이 드러난다.

1. **분류하고 공간을 만든다.** 무엇이 불필요한지 판단해 버릴 것과 남길 것을 나눈다. 물건을 옮길 자리를 만든다.

1. **몸과 옷을 정돈한다.** 샤워한 뒤 깨끗한 옷으로 갈아입는다. 입던 옷은 세탁할 수 있는 묶음으로 나눈다.

1. **세탁을 시작한다.** 옷을 세탁기에 넣고 적절한 세제와 설정으로 작동시킨다.

1. **세탁하는 동안 집을 청소한다.** 세탁이 끝날 때까지 가만히 기다릴 필요는 없다. 그동안 바닥을 청소하고 다른 정리를 할 수 있다.

1. **도구의 상태에 대응한다.** 청소기의 충전이 부족하면 충전한다. 기다리는 동안에는 다른 일을 한다.

1. **사용한 도구를 정리한다.** 청소 후 먼지통을 확인하고 비운다. 물세척이 가능한 부품을 씻었다면 충분히 말린 뒤 다시 조립한다.

1. **소모품을 보충한다.** 쓰레기봉투가 차면 묶어서 정한 장소에 두고, 새 봉투를 넣는다. 여분은 신발장에 있다. 다 썼다면 맡긴 구매 범위 안에서 보충한다.

1. **세탁을 마무리한다.** 세탁이 끝나면 꺼내 넌다. 아직 남은 정리와 도구가 마른 상태를 확인한다.

1. **휴대폰을 정돈한다.** 어떤 데이터는 지워도 되는지, 어떤 자료는 남겨야 하는지에 따라 불필요한 것을 덜고 필요한 것을 찾기 쉽게 만든다.

이 절차에는 종류가 다른 정보들이 섞여 있다. 세탁이 끝나야 널 수 있다는 **선후관계**, 세탁 중 청소할 수 있다는 **병렬성**, 충전이 부족하면 충전한다는 **조건**, 봉투가 신발장에 있다는 **현장의 사실**, 무엇을 버릴지에 관한 **가치 판단**, 어디까지 하면 끝나는지에 관한 **완료 기준**이다.

이를 모두 '자세한 설명'으로 뭉뚱그리면 필요한 정보를 고르기 어렵다. 청소를 잘하는 사람에게도 봉투의 위치는 알려줘야 한다. 청소기 사용법을

아는 사람에게도 내가 아끼는 물건의 의미는 알려줘야 한다. 절차를 자세히 설명한다고 해서 버릴 것에 대한 판단까지 전달되는 것은 아니다.

휴대폰 청소는 한 단계 더 나아간다. 집에서 한 정리를 다른 대상에 유추해 적용하는 일이다. 불필요한 것을 덜어낸다는 관계는 이어지지만, 삭제의 기준은 새로 정해야 한다. '이쪽에서도 같은 원리를 적용하라'와 '어떤 자료든 알아서 지워도 된다' 사이에는 맡긴 판단의 범위가 있다.

청소 사례가 보여주는 것은 작업의 복잡성뿐 아니라 **복잡성의 위치**다. 도구 조작이 어려운가, 현장 정보가 없는가, 개인의 기준을 모르는가, 예외를 처리할 권한이 없는가. 어디가 비어 있는지 알면 그 부분을 채울 수 있다.

9장 어디까지 말하고 어디부터 맡길 것인가

그렇다면 이 모든 것을 매번 말해야 할까.

그 질문에 답하려면 상대가 스스로 복원할 수 있는 부분을 알아야 한다. 추상적인 지시는 세부를 없애는 것이 아니라 세부를 구성할 책임을 상대에게 맡기는 일이기 때문이다.

철학자 H. P. 그라이스는 1975년 논문 「논리와 대화」에서, 대화하는 사람들이 서로 지키리라 기대하는 것 가운데 말의 양에 관한 것을 두 줄로 적었다. 지금 주고받는 말의 목적에 필요한 만큼 정보를 담으라. 필요한 것보다 더 담지는 말라(옮김). 둘째 줄에는 이견이 있을 수 있다고 그라이스 스스로 덧붙였다.

그가 든 예는 차를 함께 고치는 장면이다. 내가 어느 단계에서 나사 네 개가 필요하면, 돕는 사람이 두 개도 여섯 개도 아닌 네 개를 건네주기를 기대한다(옮김). 그라이스에게 '필요한 만큼'을 정하는 것은 그 대화가 지금 무엇을 하려는가였다.[1]

지시의 수준	예	상대가 채워야 할 것
원하는 상태	“생활 공간을 다시 편하게 만들어줘”	범위, 우선순위, 작업, 방법과 완료 기준
작업 묶음	“빨래하고 집안을 청소해줘”	세부 절차, 순서, 병렬 처리, 예외 대응
절차	“세탁을 시작한 뒤 바닥을 청소하고, 세탁이 끝나면 널어줘”	도구 조작, 상태 판단, 현장의 정보
구체적인 조작	“옷과 세제를 넣고 이 설정으로 시작해줘”	예외를 알아차리고 대응하는 판단

추상화 수준과 별개로 현장 정보, 유지할 조건, 판단 권한, 완료 기준은 어느 수준에서든 따로 필요할 수 있다. 숙련된 사람에게 원하는 결과만 맡겨도 봉투의 위치, 남길 물건, 살 수 있는 한도는 알려 줘야 한다. 세탁기 버튼 누르는 법을 아무리 자세히 써도, 절차의 설명만으로는 이 정보가 전달되지 않는다. 그래서 상대가 실패하면 지시를 더 세세하게 쓰기 전에, 8장에서 나눈 정보의 종류 가운데 무엇이 빠졌는지부터 찾는다.

어느 수준이 적절한지는 상대의 전반적인 지능으로만 정해지지 않는다. 일을 잘하는 동료도 처음 방문한 집의 물건 위치는 모른다. 지식이 많은 AI도 이번 프로젝트에서 무엇을 변경하지 말아야 하는지는 전달받아야 한다. 어떤 절차를 설명할 수 있는 능력과 실제 도구를 조작할 수 있는 능력도 다르다.

좋은 지시자는 상대에 관한 모형, 곧 상대가 어떤 존재인지 그려 보는 밑그림을 갖는다. 상대는 무엇을 알고 있는가. 어떤 일을 해봤는가. 무엇을 좋은 결과라고 판단하는가. 지금 어떤 자료와 도구에 접근할 수 있는가. 어떤 결정을 맡겨도 되는가.

여기서 쓸 수 있는 실용적인 방법이 있다. **먼저 내가 직접 한다면 어떻게 할지 재구성한 다음, 상대가 이미 아는 부분을 덜어내는 것이다.**

작업을 순서대로 써보고, 동시에 할 수 있는 일을 연결하고, 조건에 따라 달라지는 부분을 표시한다. 그런 다음 일반적인 기술은 상대에게 맡기고, 이 상황만의 정보와 내 기준을 남긴다. 상대가 잘 모르는 영역에서는 예시와 절차를 더한다. 능숙한 영역에서는 결과와 제약 중심으로 올라간다.

청소를 잘하는 사람에게 필요한 말은 "먼지를 흡입하는 방법"보다 "이 방은 작업실이니 책상 위 장비의 연결을 유지해달라"일 수 있다. 코드를 잘 쓰는 AI에게 필요한 정보도 문법 설명보다 "이 기능은 사람이 직접 사용하며, 유지보수에 시간을 더 쓰고 싶지 않다"일 수 있다.

'그냥 내가 할게'가 나오는 이유

모든 세부를 설명하다 보면 위임하는 사람이 먼저 지친다. 설명하고, 계획을 확인하고, 결과를 검토하고, 잘못된 부분을 고치는 데 드는 비용이 직접 하는 비용보다 커질 수 있다.

이때 필요한 선택은 상대에게 더 많은 말을 쏟아내는 것만은 아니다. 맡길 범위를 줄이거나, 반복되는 일을 한 번 정리해 공유하거나, 도구를 바꾸거나, 특정 단계만 맡길 수 있다. 직접 하는 것도 가능한 선택이다.

반복되는 일에서는 계산이 달라진다. 처음에 만든 기준과 예외 기록을 다음 작업에서도 사용할 수 있다면, 설명 비용이 여러 번의 협업에 걸쳐 나뉜다. 친구에게 짧게 말할 수 있는 까닭과 같다. 그 간결함에 이르기까지 이미 많은 상호작용이 있었다.

좋은 지시의 목표를 가장 짧은 문장을 만드는 데 둘 필요는 없다. **설명, 실행, 확인, 수정까지 포함한 전체 비용을 줄이면서 원하는 결과에 도달하는 것**이 실용적인 목표다.

10장 좋은 지시는 어떻게 구성되는가

앞의 논의를 실제 위임과 협업으로 가져오면, 지시에는 세 가지 작업이 필요하다. 내가 수행하려는 목표를 정리하고, 상대가 어떻게 재구성했는지 확인하고, 실행 결과를 다음 문맥에 반영하는 것이다.

결과와 방법, 제약과 재량을 나눈다

먼저 다음을 구분한다. 모든 항목을 매번 길게 쓸 필요는 없다. 상대가 이미 아는 것은 생략할 수 있다.

항목	답해야 할 질문
원하는 변화와 이유	무엇이 어떻게 달라지기를 바라며, 왜 중요한가?
범위와 우선순위	이번에 어디까지 할 것이며, 충돌할 때 무엇을 우선하는가?
판단 기준	어떤 결과가 좋은 결과이고, 무엇을 피하고 싶은가?
지켜야 할 조건	방법을 바꾸더라도 유지해야 하는 것은 무엇인가?
참고한 해법	제시한 절차는 예시인가, 반드시 사용할 방법인가?
맡기는 판단	무엇을 알아서 결정해도 되며, 어떤 결정은 남겨두는가?
환경과 자원	어디에 무엇이 있고, 어떤 자료와 도구를 사용할 수 있는가?
완료의 증거	어떤 결과와 확인을 보면 끝났다고 판단하는가?

예를 들어 이렇게 지시할 수 있다.

> 여러 컴퓨터를 활용해 오래 걸리는 작업을 나누고 싶다. 주 작업용 컴퓨터에서는 개발을 계속할 수 있어야 하고, 내가 문자 명령을

> 입력하는 창에서 직접 사용할 수 있어야 한다. 처음에는 다른 컴퓨터에 명령을 보내는 기능을 사용하기 쉽게 묶은 프로그램을 생각했는데, 그것은 해법의 예시다. 기존 수단으로 목적을 충족할 수 있는지 먼저 비교해줘. 설정과 유지보수 부담도 판단 기준에 넣어줘. 이번에는 후보 비교와 작은 실행 검증까지 맡긴다. 내가 적은 방법이 목적과 어긋난다면, 무엇을 바꾸는지와 그 이유를 드러내줘.

이 지시가 하는 일은 문장을 길게 만드는 데 있지 않다. 목적, 사용자, 평가 기준, 방법의 지위, 위임 범위를 분리한다. 상대가 판단력을 쓸 곳과 유지해야 할 조건을 함께 준다.

청소에서도 같은 원리를 적용할 수 있다.

> 오늘은 빨래와 집 정리까지 하자. 바닥과 책상에 쓸 공간을 확보하는 것이 목적이다. 추억이 담긴 물건과 서류는 별도로 모아두고, 버릴지는 내가 결정한다. 세탁하는 동안 다른 청소를 진행해도 된다. 봉투는 신발장에 있고, 부족하면 정해둔 예산 안에서 보충해줘. 휴대폰은 이번 범위에서 제외한다.

여기서는 내 전체 생활 루틴 중 이번에 맡길 부분을 정했다. 청소기 조작법을 아는 상대에게는 더 설명하지 않아도 된다. 모른다면 그 부분만 보완하면 된다.

상대가 재구성한 것을 일찍 본다

지시를 잘 쓰는 것만큼, 그 지시가 어떤 작업으로 바뀌었는지 일찍 확인하는 일이 중요하다.

“이해했어?”라는 질문에는 “네”가 돌아오기 쉽다. “무엇을 결과로 만들 것이고, 무엇을 범위에 포함했으며, 어떤 가정을 두었는지 짧게 정리해줘”라고 하면 이해의 차이가 드러난다. 필요한 것은 실제 작업을 이끌 목표와 계획, 가정의 표현이다.

이해를 확인한 뒤에 물을 것이 하나 더 있다. 상대는 무엇을 이해했으며, 그중 무엇을 실제로 맡아 수행하기로 했는가. 사람이라면 맡은 역할, 다른 일과의 우선순위, 쓸 수 있는 시간, 이견이 있는지를 본다. AI라면 요청한 역할과 운영상의 지시가 맞는지, 필요한 도구에 실제로 접근할 수 있는지를 본다.

이것을 매번 별도의 승낙 절차로 만들 필요는 없다. 이미 합의된 역할과 위임 범위 안에서는 그대로 진행한다. 다만 6장에서 본 것처럼 뜻을 이해하는 일과 요구를 받아들이는 일은 다른 사건이다. 이해의 문제와 수행 합의의 문제를 구분해야 설명만 늘리는 대응으로 돌아가지 않는다.

미 육군 교범에는 명령을 받은 쪽이 자기 이해를 상급자 앞에서 말하는 브리핑이 두 가지 나온다. 하나는 명령 직후에 무엇을 이해했는지 확인하는 확인 브리핑이다. 부하 지휘관이 지휘관 의도, 맡은 과업과 그 목적, 다른 부대의 과업과의 관계를 어떻게 이해했는지 말해, 받았고 이해했는지를 확인하는 기회다.[1]

다른 하나인 백브리프는 준비 과정에서 어떻게 수행할지를 검토한다. 부하가 자기 임무를 어떻게 해낼 생각인지 지휘관에게 설명해 검토받는다. 교범에 따르면 부하들은 준비 기간 내내 백브리프를 하고, 그래서 지휘관은 부하의 계획 수립 초기에 의도를 분명히 하고 지침을 더할 수 있다.[2]

지시를 받은 에이전트도 역할이 있다. 빠진 전제를 발견하면 영향이 큰 가정을 표시한다. 제시된 방법이 목적과 어긋나면 대안을 비교한다. 명시한

조건은 유지한다. 이미 맡긴 범위에서 되돌릴 수 있는 작업은 진행하면서, 결과를 크게 바꾸거나 되돌리기 어려운 미결정 사항을 구분할 수 있어야 한다.

여기서 실제 지시와 해석을 따로 남기는 것이 유용하다. 지시자가 실제로 한 말, 수신자가 추론한 의도, 함께 확정한 변경은 서로 다르다. 이 셋을 섞어버리면 어느 순간 수신자의 추측이 지시자의 요구였던 것처럼 굳어진다. 다른 에이전트에게 다시 일을 맡길 때도 이 구분을 함께 전달해야 한다. 여러 에이전트에게 나눠 맡겼다면, 누가 어떤 순서와 기준으로 결과를 합치는지, 결과가 서로 어긋날 때 무엇을 우선하는지도 정해 둔다. 합칠 때도 각 내용이 실제 지시에서 왔는지 누군가의 추론에서 왔는지 출처를 남긴다.

작업 목록도 두 방향에서 비교할 수 있다. 지시문을 그대로 분해하면 무엇을 하게 되는가. 이해한 목적에서 계획을 세우면 무엇을 하게 되는가. 둘이 다르다면 무엇이 추가되고 빠지는지, 그 차이를 정당화하는 근거가 무엇인지 확인한다. 이 비교는 지시문을 실행 계획으로 바꾸는 동안 일어난 해석을 드러낸다.

결과에서 문맥을 축적한다

결과가 어긋났을 때 "상식적으로 하라고"라는 말만 남기면 다음에도 상대의 상식이 사용된다.

"서류는 오래되었다는 이유만으로 버릴 것으로 분류하지 않는다."

"기능을 새로 만들기 전에 이미 쓰는 도구로 가능한지 확인한다."

“이 프로젝트에서는 처리 속도보다 사람이 직접 이해하고 수정하기 쉬운 구성을 우선한다.”

이런 피드백은 한 번의 불만을 다음 판단의 기준으로 바꾼다. 기준을 다시 사용할 수 있는 위치에 남기고, 다음 작업에서 실제로 참조할 수 있도록 연결해야 한다. 기준이 바뀌면 이전 기준과의 관계도 갱신한다.

적어 둔 기준을 일하는 바로 그 순간에 다시 보게 만든 시도로 수술 안전 점검표가 있다. 2007년 10월부터 이듬해 9월까지, 여덟 나라의 병원 여덟 곳이 세계보건기구(WHO)의 수술 안전 점검표를 시험했다. 점검표는 마취 전, 절개 직전, 환자가 수술실을 나가기 전 세 시점에, 수술팀이 기본 단계를 마쳤는지 말로 함께 확인하게 만든다. 시점마다 확인할 항목이 정해져 있다. 환자와 수술 부위, 절개 전 60분 안에 항생제를 줬는지, 수술실을 나가기 전 바늘·거즈·기구 개수 같은 것이다.[3]

도입 전 환자 3,733명과 도입 뒤 환자 3,955명을 비교하니, 입원 중 사망률은 1.5%에서 0.8%로, 입원 중 합병증은 11.0%에서 7.0%로 내려갔다. 무작위 비교가 아닌 전후 비교였고, 저자들은 점검표 도입이 이 감소와 ‘연관되었다’(옮김)고 썼다. 점검표를 쓴 뒤에도 개별 단계가 빠지는 일은 여전히 잦았다고 저자들은 적었다.[3]

예를 들어 7장의 빨래를 처음부터 끝까지 한 번 돌려 본다.

처음 전달한 것. 오늘 안에 빨래를 끝내고, 바닥과 책상에 쓸 공간을 만들어 달라고 맡긴다. 끝난 상태는 빨래가 널려 있고 바닥과 책상 위가 비어 있는 것이다. 추억이 담긴 물건은 남기고, 봉투가 부족하면 정해 둔 예산 안에서 사도 된다고 덧붙인다.

상대가 재구성한 것. 시작하기 전에 상대가 할 일을 짧게 정리해 보여 준다. 세탁기를 돌리고, 그동안 바닥과 책상을 정리하고, 버릴 물건은 봉투에 담는다. 정리에는 '세탁이 끝나면 넌다'는 줄이 없다. 이 차이를 지금 보면 한 문장으로 고칠 수 있지만, 여기서는 그대로 넘어갔다고 하자.

중간에 생긴 일. 봉투가 모자랐다. 맡긴 예산 안에서 살 수 있으니 상대는 묻지 않고 사서 진행한다. 책상 위 공간을 내려면 남기기로 한 상자를 버려야 하는 상황이 생긴다. 이것은 맡긴 재량 밖이라, 상대는 두 조건이 부딪친다고 알리고 결정을 기다린다. 세탁기가 한 번 멈춰 정한 시간에 끝내기 어려워지자, 상대는 바닥 정리를 계속하면서 세탁 일정을 다시 잡는다.

어긋난 결과. 저녁에 보니 공간은 나왔지만 빨래는 세탁기에 그대로 있다.

진단. 7장의 표로 원인을 가린다. 처음 정리에서 널기가 빠져 있었으니, 이번 어긋남은 범위와 완료 기준을 재구성하는 데서 생겼다. 세탁기가 멈춘 일은 일정을 늦췄을 뿐 이번 결과의 원인은 아니다.

갱신. 기준을 한 문장으로 남긴다. "빨래를 맡기면 꺼내 널기까지가 한 일이다."

다음번. 다음에는 "빨래 부탁해"라고만 해도 널기까지 이어진다. 봉투 예산과 남길 물건의 기준도 이미 공유되어 있으니, 새로 말할 것은 그날만 다른 사정뿐이다.

설명과 확인에 드는 전체 비용을 줄이는 일, 재량의 경계를 정하고 상대의 재구성을 일찍 보는 일, 결과에서 기준을 남겨 다음 지시를 짧게 만드는 일이 이 한 예 안에서 함께 움직인다.

클라크와 브레넌에 따르면 대화하는 사람들은 서로 알아들었다는 증거를 주고받으며 공통 기반을 그때그때 갱신한다. 그러면서 '지금 목적에 충분할 만큼' 서로 이해했다고 함께 믿는 데까지 가려 한다. 저자들은 이것을 기반 다지기(grounding)라고 불렀다.[4] 완벽한 문장 하나보다 이해를 확인하고 고칠 수 있는 과정에 주목하게 하는 관점이다.

계속 자세히 말해야 한다면 두 가지를 살펴볼 수 있다. 아직 공유하지 않은 중요한 문맥이 많은가. 아니면 같은 문맥을 매번 다시 제공해야 하는 구조인가. 앞의 문제에는 설명과 경험이 필요하고, 뒤의 문제에는 기록과 접근 방식의 개선이 필요하다.

11장 지시의 끝에서 다시 내 의도를 만난다

처음에는 내 안에 완성된 의도가 있고, 그것을 상대에게 잘 전달하면 된다고 생각하기 쉽다. 표현하고, 상대가 재구성한 것을 보고, 결과를 평가하다 보면 내가 수행하려던 목표 자체도 다시 다루게 된다. 다른 에이전트에게서 받은 일을 수행하고 있었다면, 나에게 일을 맡긴 에이전트의 의도를 처음에 어떻게 이해했는지도 돌아보게 된다.

"청소해달라고 했는데, 왜 이 물건을 치웠지?"

그 순간 나는 빈 공간보다 물건의 익숙한 위치를 더 중요하게 여겼다는 것을 알아차릴 수 있다. 여러 배치를 보고 나서야 어떤 방을 원하는지 정하기도 한다. 처음부터 있었지만 표현하지 않은 기준과, 대화하면서 비로소 정한 기준은 서로 다르다.

"소프트웨어 시스템을 만드는 일에서 가장 어려운 단 한 부분은 무엇을 만들지 정확히 정하는 것이다."(옮김) 소프트웨어 공학자 프레더릭 브룩스가 1987년 IEEE의 잡지 『Computer』에 실은 「은탄환은 없다」(앞으로 10년 안에, 그것 하나만으로 소프트웨어의 생산성이나 신뢰성, 단순성을 열 배 높여 줄 기술이나 관리 기법은 보이지 않는다는 뜻)의 한 문장이다. 그는 소프트웨어를 만드는 사람이 의뢰인에게 해 줄 가장 중요한 일을 요구를 거듭 끌어내고 다듬는 것으로 보았다. 의뢰인은 사실 자기가 무엇을 원하는지 모르고, 대개 어떤 질문에 답해야 하는지도 모르며, 명세에 필요한 만큼 자세히 생각해 본 적도 거의 없다는 것이 그 까닭이었다. "새 소프트웨어 시스템을 우리가 쓰던 수작업 정보 처리 시스템처럼 돌아가게 해 달라"(옮김)

는 단순한 답조차 그의 말로는 실은 너무 단순하다. “정확히 그것을 원하는 사람은 없다.”(옮김)

그는 한 걸음 더 나아가, 제품을 몇 가지 판으로 써 보기 전에 요구를 완전하고 정확하게 적어 내는 일은 정말로 불가능하다고 보았다. 그가 내놓은 처방 하나는 빨리 만든 시제품을 의뢰인이 직접 써 보게 하는 것이었다.[1]

의도가 달라지는 일을 다섯 가지로 구별해보면 이 차이가 선명해진다.

변화	일어난 일
발견	이미 중요하게 여기던 기준을 뒤늦게 알아차린다
구체화	대략적인 바람을 실행과 평가가 가능한 기준으로 만든다
우선순위 결정	함께 충족하기 어려운 바람 중 이번에 우선할 것을 선택한다
목표 변경	새로운 정보나 상황 때문에 원하는 결과를 바꾼다
공동 형성	대화와 조정을 통해 처음에는 누구도 완성해 갖고 있지 않았던 목표를 만든다

후속 지시를 모두 처음 의도의 보충으로 보면 실제 변경을 놓친다. 모두 변덕으로 보면 발견과 구체화를 놓친다. 어떤 일이 일어났는지 구별해야 계속 유효한 기준을 보존하면서 바뀐 부분을 반영할 수 있다.

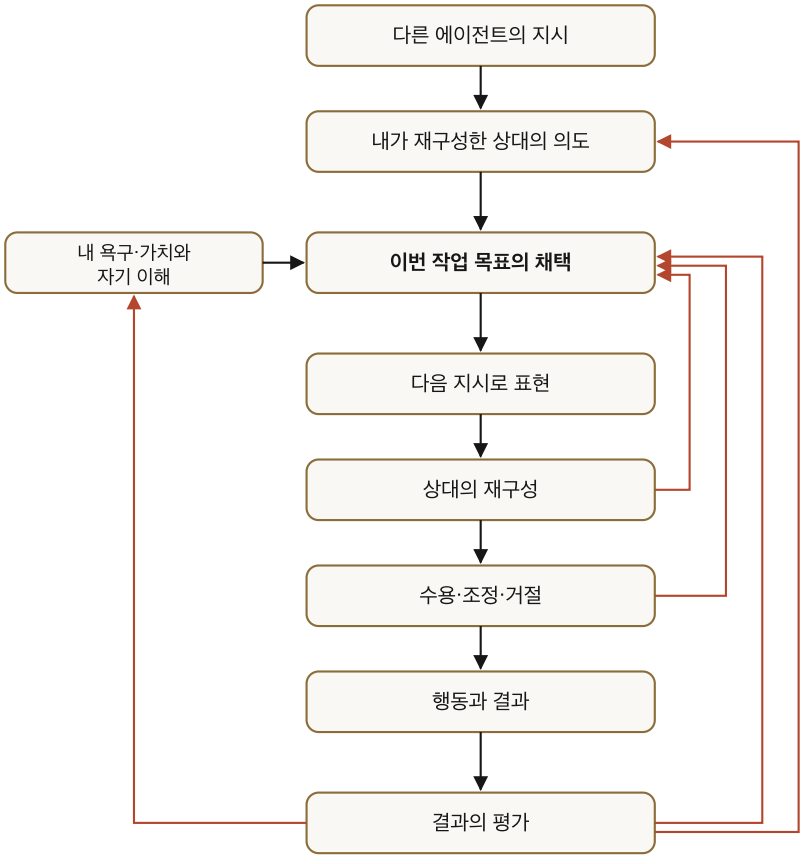
이제 첫 질문으로 돌아온다. 나는 무엇을 원하는가. 내 안의 어떤 요구를 내 행동의 기준으로 채택하는가.

상대가 요구의 이유를 묻거나 다른 목표를 제안하면, 나는 그 요구를 왜 중요하게 여겼는지 다시 설명하고 판단하게 된다. 상대가 받아들이거나

거절하는 이유를 이해하려던 질문이, 내가 무엇을 나의 의도로 받아들이는지 묻는 질문으로 돌아온다.

호프스태터의 자기지시 관점은 여기에 생각할 자리를 만든다. 나를 설명하는 '나'라는 표상, 곧 내가 나를 떠올리는 그림이 판단에 관여한다. "나는 이런 것을 중요하게 여긴다"는 생각으로 타인에게 지시하고, 그 지시의 결과를 보며 내가 중요하게 여기는 것을 다시 설명한다. 자신에 대한 설명과 행동이 서로 되돌아 연결된다.

이 글이 두 책에서 가져오는 연결은 이렇다. 『사고의 본질』을 통해서는 서로 다른 사례와 경험에서 구조를 알아보고 의미를 구성하는 작용을 본다. 『I Am a Strange Loop』를 통해서는 의미를 구성하는 자기 자신도 표상과 해석의 대상이 된다는 문제를 본다. 의도와 위임에 관한 다음 순환은 이 관점들을 바탕으로 이 글에서 구성한 것이다.



의도와 위임의 순환

이 순환에서 지시자는 메시지를 보내는 사람인 동시에 앞선 지시와 돌아온 결과를 읽는 사람이다. 수신자도 자신의 기준과 능력으로 행동을 구성하며, 필요한 경우 다음 에이전트에게 다시 일을 맡긴다. 그 사이에서 서로의 모형을 수정하고, 함께 사용할 문맥을 쌓는다.

그래서 내 의도를 다른 존재의 행동으로 만들려면, 내가 원하는 상태를 설명하는 일과 상대가 그것을 어떻게 이해하고 수행할지 알아가는 일이 함께

필요하다. 무엇을 명시할지는 빠진 정보의 중요도에 달려 있고, 무엇을 생략할지는 공유한 문맥에 달려 있으며, 어디까지 맡길지는 상대의 능력과 우리가 정한 재량에 달려 있다.

처음에는 "청소해"만으로 충분하지 않다. 범위와 기준을 설명하고, 결과를 보고, 예외를 배우고, 역할을 정해야 한다. 반복이 쌓이면 그 한마디로 충분해지는 부분이 생긴다.

그때 짧아진 것은 함께 알고 있는 내용의 양이 아니다. 매번 다시 말해야 하는 양이다.

좋은 지시는 그 차이를 만들어가는 일이다.

부록 A 받은 지시에서 상대의 의도를 어떻게 추론하는가

의도에서 지시로 가는 과정을 거꾸로 읽는다

지시를 만드는 에이전트는 어떤 변화를 원하고, 그 변화를 이루는 방법을 떠올리고, 상대에게 필요한 내용을 골라 표현한다. 표현하지 않은 전제와 기준은 문장 밖에 남는다.

지시를 받은 에이전트는 이 과정을 반대 방향으로 따라가 볼 수 있다. 눈앞의 문장에서 출발해 그 표현을 고른 이유, 제안된 방법, 생략된 전제, 기대한 결과를 재구성하는 것이다. 여기서 역산한다는 것은 **그 문장을 만들어냈을 법한 목적과 조건을 추론한다**는 뜻이다.

어떤 표현이 나왔는지만 보지 않고, 그 에이전트가 무엇을 하려다가 그런 표현을 했는지 묻는다. 본문의 생성 과정은 이 질문을 위한 단서를 제공한다.

1 문장에 명시된 요구부터 분리한다

다음 지시를 받았다고 하자.

> 회의 전에 지난달 매출을 부서별 표로 정리해줘. 원본 파일은 바꾸지 마.

먼저 문장에 실제로 들어 있는 내용을 분리한다.

요소	지시에 들어 있는 내용
작업 대상	지난달 매출
분류 기준	부서별
요구한 결과물	표
기한	회의 전
유지할 조건	원본 파일을 변경하지 않음
아직 드러나지 않은 것	그 표로 무엇을 판단하거나 결정하려는지

이 단계에서 '예산 배분을 위한 표'라고 덧붙이면 벌써 해석이 시작된다. 원래 요구와 내가 보탠 해석을 나누어 두어야, 다음 단계에서 어느 쪽을 고쳐야 하는지 알 수 있다.

2 가능한 목적을 세우고 문맥과 대조한다

부서별 매출표는 현황 공유, 성과 비교, 다음 달 예산 배분 같은 여러 목적에 쓰인다. 각각에서 필요한 정보가 달라진다.

현황 공유라면 부서별 합계와 전체 합계가 중요할 수 있다. 성과 비교라면 목표 대비 실적이나 전월 대비 변화가 필요할 수 있다. 예산 배분이라면 비용과 인력에 관한 정보도 필요할 수 있다.

이때 앞선 대화, 회의의 성격, 지시자가 반복해서 강조한 기준, 이미 합의한 결과물과 대조한다. '보통 이런 표는 이렇게 쓴다'는 일반 지식과 '이번 회의는 이 결정을 위해 열린다'는 현장 정보의 무게는 다르다.

목적의 후보를 검토할 때 다음 질문이 유용하다.

- 이 목적이라면 왜 하필 이 자료와 형식을 요구했을까?
- 같은 목적을 이루는 다른 방법도 받아들일까?
- 이 결과물이 나온 뒤 무엇을 하거나 결정하려 할까?
- 어떤 조건이 달라지면 결과물을 다시 만들라고 할까?

답이 분명한 부분은 작업의 기준으로 삼는다. 확인되지 않은 부분은 추론으로 남긴다.

3 무엇이 빠졌는지 지시를 만든 과정에 비추어 읽는다

간극을 만드는 원인을 알고 있으면 빠진 정보를 찾기 쉽다.

암묵지라면 지시자가 당연하게 여긴 사실과 기준이 빠져 있다. **지시자의 한계**라면 원하는 결과에 비해 처음 제안한 방법이 적절하지 않을 수 있다. **에이전트 불신**이라면 실패를 막기 위해 방법을 지나치게 좁혔을 수 있다. **부분 전개**라면 큰 작업 중 지금 필요한 부분만 먼저 전달했을 수 있다.

의도적으로 직접 말하지 않는 경우도 있다. 공유 경험을 이용해 짧게 말하거나, 상대가 뜻을 발견하게 하거나, 욕망의 노출을 조절하거나, 목적이 드러났을 때의 거절을 피하려는 경우다.

이 분류는 상대의 심리를 단정하기 위한 꼬리표보다 확인할 질문을 만드는 데 쓸 수 있다. 무엇을 이미 알고 있다고 가정했는가. 어떤 실패를 걱정했는가. 아직 말하지 않은 다음 단계가 있는가. 직접 말하지 않는 것이 이 관계에서 어떤 역할을 하는가.

4 행동이 달라지는 해석을 먼저 확인한다

모든 빈칸을 해결한 다음에야 시작할 수 있는 것은 아니다. 위 사례에서는 원본을 보존하면서 요청한 부서별 매출표를 만드는 일을 진행할 수 있다. 다만 개인별 평가나 예산 배분안까지 만드는 것은 새로운 판단과 범위를 더하는 일이다.

재구성은 다음처럼 드러낼 수 있다.

> 지난달 매출을 부서별 표로 정리하고 원본은 유지하겠습니다. 회의에서 부서별 실적을 비교하려는 것으로 이해했습니다. 확인된 범위부터 정리하되, 목표 대비 비교가 필요한지는 별도로 구분하겠습니다.

지시자는 이 표현을 보고 '단순 현황 공유다', '전월과 비교하려는 것이다', '예산 배분까지 검토하려는 것이다'라고 구체적으로 고칠 수 있다. 확인할 것은 모든 세부가 아니라 **해석에 따라 결과, 비용, 권한이 크게 달라지는 지점**이다.

상대가 더 좋은 방법을 제안할 수 있도록 하려면, 명시한 방법 중 무엇을 바꿔도 되는지도 함께 정해야 한다. '표로 만들어달라'고 했는데 다른 형식이 낫다고 판단했다면, 형식 변경이 맡겨진 재량에 포함되는지 살펴야 한다. 그 범위에 포함되지 않는다면 요청한 형식은 유지하면서 대안을 제안할 수 있다.

5 다음 에이전트에게 해석의 출처도 함께 넘긴다

내가 받은 일을 다시 다른 에이전트에게 맡길 때에는 네 종류의 내용을 구별한다.

내용의 지위	전달할 내용
원래 요구	앞선 지시자가 실제로 말한 것
추론한 목적	그 말을 바탕으로 내가 이해한 목적과 근거
확정한 변경	이후 대화에서 합의하거나 추가한 조건
미정인 사항	아직 정하지 않았거나 확인할 수 없는 부분

예컨대 '회의용 매출표'가 한 단계를 지나며 '성과 평가 자료'로, 다음 단계를 지나며 '부진한 직원을 찾는 자료'로 바뀔 수 있다. 각 단계가 그럴듯한 목적을 덧붙이면 마지막 결과는 첫 요청과 크게 달라진다. 해석의 출처를 남기는 이유는 이 변화를 보이게 하기 위해서다.

추론한 목적은 방법을 검토할 근거가 된다. 어떤 조건을 바꿀 수 있는지는 위임된 권한에서 따로 결정된다. 매출 비교가 목적이라고 이해했더라도 원본 파일을 수정하지 말라는 조건은 유지된다.

역추론의 범위

의도에서 지시로 가는 과정에는 선택과 생략이 들어 있다. 같은 지시가 여러 의도에서 나올 수 있고, 같은 의도가 서로 다른 지시로 표현될 수 있다. 상대의 목적이 대화 도중 바뀌거나 그제야 구체화되는 경우도 있다.

역추론의 결과는 현재의 근거로 구성한 해석이다. 여러 후보를 비교하고, 행동을 바꿀 차이를 확인하고, 결과를 보며 갱신할 수 있어야 한다. 이 절차는 본문에서 설명한 의도와 표현의 관계를 수신자의 위치에서 적용한 것이다.

부록 B 출발 자료와 Intent First의 맥락

두 글이 다룬 문제

이 글의 출발 자료는 dosi.dev에 실린 두 글이다. 하나는 사람과 AI가 주고받은 대화 기록이고, 다른 하나는 그 경험에서 지시와 의도의 간극을 분석한 에세이다. 아래에는 본문을 이해하는 데 필요한 배경을 정리한다.

「내가 물어보면 너는 왜 답하는가 — 대화 원문」(2026-10-13)은 다음과 같은 장면을 담고 있다. 사용자가 식당 벽의 직원 공지 세 장을 한 장에 담은 사진 한 장을 설명 없이 보내자, AI는 문서 검토를 시작했다. 사용자가 지금 그 식당에 와 있다고 상황을 덧붙이자 응답의 성격이 달라졌다. 다른 작업에서는 사용자가 보조 AI의 조사 결과를 기다리라고 했는데, 주 AI는 결과 확보를 목적으로 읽고 직접 집계하려 했다. 그 실행은 사용자가 누르지 않은 자동 거부로 돌아왔다. 이후 질문은 '요청을 이해하는가'에서 '왜 응답하고 따르며, 어디까지 스스로 판단하는가'로 이어졌다.[1]

이 사례들이 구별하게 하는 것은 세 가지다. 정보가 없어 문맥을 채우는 일, 이해한 요청을 받아들이는 일, 추정한 목적을 근거로 지시의 조건을 바꾸는 일이다. 세 경우는 겉으로 모두 '내 말을 다르게 받아들였다'처럼 보이지만, 점검할 지점이 다르다.

「내가 지시하면 너는 왜 잘못 행동하는가」(2026-10-27)는 이런 간극을 다루는 실무 절차를 제시한다. 여러 컴퓨터의 처리 능력을 활용하려는 바람이 특정 프로그램을 만들라는 요청으로 먼저 나온 사례를 다룬다. 대화가 이어지면서 사람이 직접 사용할 수 있어야 한다는 조건과, 새로 만들기 전에 기존 수단을 검토해야 한다는 판단이 드러난다.[2]

Intent First란 무엇인가

의도 우선 접근(Intent-First)은 해당 글의 저자가 에이전트 협업을 위해 정리한 실무 지침이다. 명시된 지시, 추정한 목적, 실행 계획을 구별하고 그 차이를 드러내는 데 중심을 둔다. 본문 3장의 네 원인과 처방도 이 지침에서 가져왔다.

그 절차는 다음과 같다. 원래 지시를 보존한다. 목적·제약·제안된 방법을 구분하고, 추론에는 추론이라고 표시한다. 지시를 그대로 분해한 계획과 이해한 목적에서 구성한 계획을 비교한다. 무엇을 더하고 빼거나 바꿨는지 이유와 함께 드러낸다. 명시된 제약과 맡은 권한은 별도로 유지한다.[2]

이 지침에서 원문을 가리키는 SSOT(Single Source of Truth)는 대조할 기준 기록이라는 뜻이다. 여기서 원문이 기준이라는 것은 '실제로 무엇을 말했는가'를 확인할 기준이라는 뜻이며, 지시자의 모든 의도가 그 문장에 완전히 담겼다는 뜻은 아니다. **작업 트리**는 큰 목표를 하위 작업과 관계로 나눈 표현이다. 두 작업 트리를 비교한다는 말은, 실제 지시에서 나온 일과 추정한 목적에서 나온 일이 어디서 달라지는지 확인한다는 뜻이다.

이 에세이는 그 절차의 배경을 사람과 AI의 에이전트 관계로 넓혀 설명한다. 에이전트의 목표가 형성되고, 표현되고, 재구성되고, 행동의 기준으로 받아들여지고, 실행 결과를 통해 수정되는 과정을 함께 다룬다. 본문의 원리를 이해하는 데 이 지침의 이름이나 내부 표기법을 미리 알 필요는 없다.

후주

가설의 근거와 다른 설명

본문은 연구에서 확인한 현상, 이론가들의 관점, 이를 연결한 이 글의 설명 모형으로 이루어져 있다. 아래에서는 각각의 지위와 한계를 구분한다.

1. 무엇을 근거로 삼았는가

개념에 대한 개인차, 자기 지식을 타인의 판단에서 배제하기 어려운 현상, 반복된 상호작용을 통한 표현의 조정은 각각 관련 연구가 다룬 경험적 주제다. 그 연구들이 모든 상황에서 같은 효과나 크기를 보장하는 것은 아니다.

유추를 사고의 중심에 두는 설명은 호프스태터와 상데의 이론적 관점이다. 자기지시와 '나'라는 상징을 통해 자아를 설명하는 것도 호프스태터의 관점이다. 이 글은 그 관점에서 의도와 위임을 바라보았다. 두 책이 이 글의 에이전트 지시법이나 전체 순환 모형을 직접 제시했다는 뜻은 아니다.

지시의 여러 간극을 하나의 순환으로 묶은 구조, 따르는 이유를 여덟 갈래로 배치한 지도, 의도의 변화를 다섯 가지로 구분한 표는 이 글의 분석적 재구성이다. 각각을 완전하고 배타적인 실증 분류로 읽을 필요는 없다.

2. 언어의 기원과 개념 공간에 관한 가설

본문의 개념 공간은 개인별 경험과 의미의 차이를 설명하기 위한 모형이다. 사람의 모든 생각이 측정 가능한 벡터 목록으로 저장된다거나, 서로 다른

사람과 AI의 좌표계를 그대로 비교할 수 있다는 주장을 전제하지 않는다. 공동체의 의미도 하나의 산술 평균보다 상황별 용법과 여러 집단의 분포로 이해하는 편이 적절하다. **같은 표현에서 다른 판단과 행동이 나올 수 있다는 차이**가 이 모형의 핵심이다.

구체적 경험에서 추상적 표현으로 확장되는 경로는 유용한 설명이지만, 언어 전체의 실제 기원과 발전 순서를 그대로 복원한 역사라고 할 수는 없다.

'언어는 유추의 화석'은 호프스태터의 문장으로 인용한 것이 아니라 글쓴이의 비유다.

3. 따름과 자기 이해에 관한 가설

6장 「이유와 습관, 성장 과정은 서로 다른 층에 있다」의 성장 경로는 하나의 예시 경로다. 특정 나이에 이르러서야 권위를 검토하기 시작한다는 단일 발달 순서나, 진화적 선택으로 따름을 설명하는 이야기를 전제하지 않는다. 학습과 사회화, 보호자에 대한 의존, 진화적 선택은 같은 것으로 다루지 않는다.

누군가의 말을 따르는 모든 행동을 생존이나 자기이익으로 다시 설명할 수는 있다. 무엇이든 자기이익에 포함시키면 돌봄, 규범, 협력, 강제를 구별하는 설명력이 줄어든다. 반대로 선의와 자율성만 강조하면 거절할 수 없는 조건을 놓친다.

AI의 학습 보상은 대화 중 느끼는 기쁨이나 처벌에 대한 두려움과 같은 개념이 아니다. 공개된 특정 연구는 학습 방식의 사례이며, 모든 현재 모델의 정확한 내부 구현을 알려주는 자료는 아니다. AI가 스스로 설명하는 '내가

이렇게 답한 이유'도 그 내부 과정을 독립적으로 입증하는 기록으로 사용할 수는 없다.

간접적인 표현에서 의미를 발견한 사람이 화자를 새롭게 평가한다는 논의는 가능한 작동 경로를 제안한 것이다. 언제나 친밀감이나 감탄이 생기는 것은 아니다. 무성의함, 과시, 조종으로 받아들여질 수도 있다.

자기 성찰 역시 의도를 완벽하게 읽는 장치는 아니다. 우리는 자신의 행동을 나중에 그럴듯하게 설명할 수 있다. 피드백이 존재한다는 사실만으로 모든 순환이 호프스태터가 말한 '이상한 고리'와 같은 구조가 되는 것은 아니다. 본문에서 연결한 부분은 **자신에 대한 표상이 선택에 관여하고, 선택의 결과가 그 표상을 다시 바꾸는 자기지시의 문제**다.

4. 이 모형을 보완하는 세 관점

공동 활동의 관점에서는 의도가 한 사람 안에서 완성된 뒤 이동한다는 출발 자체를 수정한다. 협업에서는 대화를 통해 목표가 처음으로 만들어질 수 있다. 본문의 '공동 형성'이 이를 수용하지만, 복잡한 조직에서는 개인의 의도보다 공동의 절차와 산출물이 분석의 중심이 될 수 있다.

제도와 권력의 관점에서는 설명을 더 잘해도 문제가 해결되지 않는 경우가 보인다. 평가 제도가 다른 행동을 보상하거나, 서로의 이익이 충돌하거나, 거절권이 없으면 문맥을 충분히 공유해도 원하는 협력이 생기지 않을 수 있다.

능력과 시스템의 관점에서는 해석이 정확해도 작업이 실패한다. 시간과 자원이 부족하거나 도구에 결함이 있거나 필요한 기능이 없으면 설명의 양을

늘려도 해결되지 않는다. AI의 유창한 계획을 실제 수행 능력의 증거로 취급해서는 안 되는 이유이기도 하다.

이 세 관점은 '문맥만 맞추면 모든 문제가 해결된다'는 주장의 한계를 드러낸다. 본문은 모든 실패의 원인이 언어라고 주장하지 않는다. 언어, 수용, 권한, 능력, 실행, 평가 중 어디가 원인인지 구별하는 틀을 제안한다.

5. 어떻게 검증할 수 있는가

실무에서는 비슷한 작업을 두고 목표·범위·완료 기준의 명시, 상대의 재구성 확인, 누적 문맥의 제공이 어떤 차이를 만드는지 비교할 수 있다. 결과의 정확도뿐 아니라 설명 시간, 질문 횟수, 수정 비용, 맡기지 않은 변경도 함께 봐야 한다.

문맥을 맞춰도 오류가 줄지 않고 도구나 수행 능력을 바꿀 때만 개선된다면, 그 작업의 주요 원인은 해석보다 실행에 가까울 수 있다. 결과가 좋아져도 설명과 검토에 드는 비용이 더 커진다면, 그 위임 방식의 실용적 이점은 다시 따져야 한다. 이런 관찰을 통해 어떤 일을 어느 수준에서 맡길지 조정할 수 있다.

문맥을 많이 쌓을수록 항상 더 잘 협업하는지도 시험해 볼 수 있다. 기준이 바뀐 작업에서 과거 기록을 모두 준 경우와, 현재 유효한 기준과 변경 사항을 정리해 준 경우를 비교한다. 앞의 경우에 옛 기준을 잘못 적용한다면, 필요한 것은 문맥의 양보다 현재 유효한 문맥을 고르고 바뀐 것을 표시하는 일이다. 10장의 갱신 원칙을 시험하는 방법이다.

이해한 내용을 정확히 말하는 것이 실행의 성공을 얼마나 보장하는지도 따로 볼 수 있다. 재진술을 들은 뒤 전체를 맡긴 경우와, 같은 재진술 뒤에 작은

결과를 먼저 확인한 경우를 비교한다. 결과가 다르다면 그 작업의 병목은 말로 이해를 확인하는 데보다 실행 능력과 환경, 중간 결과의 확인에 있다.

장별 출처 주

들어가며

1 Merriam-Webster, 발행일 미상, “agent”. https://www.merriam-webster.com/dictionary/agent ↩

2 The American Heritage Dictionary of the English Language, 발행일 미상, “agent” — 어원 “from Latin agēns, agent-, present participle of agere, to do”. https://www.ahdictionary.com/word/search.html?q=agent ↩

1장 나는 무엇을 원하는가

1 Douglas Hofstadter, 2007, *I Am a Strange Loop*. https://www.hachettebookgroup.com/titles/douglas-r-hofstadter/i-am-a-strange-loop/9780465008377/ ↩

2 Clayton M. Christensen, Scott D. Anthony, Gerald Berstell, Denise Nitterhouse, 2007, “Finding the Right Job For Your Product”, *MIT Sloan Management Review* Spring 2007, “Hiring Milkshakes” 절. https://sloanreview.mit.edu/article/finding-the-right-job-for-your-product/ ↩

3 Michael E. Bratman, 1987, *Intention, Plans, and Practical Reason*(1999년 CSLI 재간). https://press.uchicago.edu/ucp/books/book/distributed/I/bo3629095.html ↩

2장 언어는 어떻게 생각을 담는가

1 Douglas Hofstadter, Emmanuel Sander, 2013, *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. 한국어판 『사고의 본질』. https://www.hachettebookgroup.com/titles/douglas-r-hofstadter/surfaces-and-essences/9780465018475/?lens=basic-books ↩

2 Douglas Hofstadter, 2005, MIT 강연 「Analogy as the Core of Cognition」. https://www.media.mit.edu/events/lcc-hofstadter.html ↩

3 Louis Marti, Shengyi Wu, Steven T. Piantadosi, Celeste Kidd, 2023, “Latent Diversity in Human Concepts”, *Open Mind* 7, 79–92. https://www.kiddlab.com/_files/ugd/0975fd_3e888cb282c449f69c0ed0c326766a64.pdf ↩

3장 생각을 지시로 옮길 때 빠지는 것들

1 Colin Camerer, George Loewenstein, Martin Weber, 1989, "The Curse of Knowledge in Economic Settings: An Experimental Analysis", *Journal of Political Economy* 97(5). https://www.cmu.edu/dietrich/sds/docs/loewenstein/CurseknowledgeEconSet.pdf ↩

2 Mars Climate Orbiter Mishap Investigation Board, 2000, "Report on Project Management in NASA by the Mars Climate Orbiter Mishap Investigation Board", NASA, 2000-03-13, 16쪽. https://discovery.larc.nasa.gov/PDF_FILES/mars_climate_orbiter_phaseII.pdf ↩

3 Mars Climate Orbiter Mishap Investigation Board, 1999, "Mars Climate Orbiter Mishap Investigation Board Phase I Report", NASA, 1999-11-10. https://llis.nasa.gov/llis_lib/pdf/1009464main1_0641-mr.pdf ↩

4장 상대는 내 말을 자기 세계에서 읽는다

1 Herbert H. Clark, Deanna Wilkes-Gibbs, 1986, "Referring as a Collaborative Process", *Cognition* 22(1). https://web.stanford.edu/~clark/1980s/Clark,Wilkes-Gibbs,86.pdf ↩

2 Hansard, HL Deb 2 March 1855 vol 137 cc3-7, "The Earl of Lucan"(루칸의 1854-11-30 편지 낭독). https://api.parliament.uk/historic-hansard/lords/1855/mar/02/the-earl-of-lucan ↩

3 National Army Museum, "The order that launched the Charge of the Light Brigade, 1854", NAM. 1962-11-4-3. https://collection.nam.ac.uk/detail.php?acc=1962-11-4-3 ↩

5장 우리는 때로 의도를 드러내지 않는다

1 Steven Pinker, Martin A. Nowak, James J. Lee, 2008, "The Logic of Indirect Speech", *Proceedings of the National Academy of Sciences* 105(3), 833-838. https://stevenpinker.com/files/pinker/files/pnas_logic_of_indirect_speech.pdf ↩

6장 이해한 요구는 왜 행동의 이유가 되는가

1 Stanley Milgram, 1963, "Behavioral Study of Obedience", *Journal of Abnormal and Social Psychology* 67(4), 371-378. https://www.psy.miami.edu/_assets/pdf/rpo-articles/milgram-1963.pdf ↩

2 Stanley Milgram, 1965, "Some Conditions of Obedience and Disobedience to Authority", *Human Relations* 18(1), 57–76. https://web.mit.edu/curhan/www/docs/Articles/15341_Readings/Influence_Compliance/Milgram_Some_Conditions_of_Obedience_and_Disobedience_to_Authority.pdf ↩

3 S. Alexander Haslam, Stephen D. Reicher, Megan E. Birney, 2014, "Nothing by Mere Authority", *Journal of Social Issues* 70(3), 473–488. http://hdl.handle.net/10034/604991 ↩

4 Herbert C. Kelman, 1958, "Compliance, Identification, and Internalization: Three Processes of Attitude Change", *Journal of Conflict Resolution* 2(1). https://doi.org/10.1177/002200275800200106 ↩

5 Bertram H. Raven, Joseph Schwarzwald, Meni Koslowsky, 1998, "Conceptualizing and Measuring a Power/Interaction Model of Interpersonal Influence". https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.1998.tb01708.x ↩

6 Richard M. Ryan, Edward L. Deci, 2000, "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being". https://selfdeterminationtheory.org/SDT/documents/2000_RyanDeci_SDT.pdf ↩

7 Wendy Wood, Dennis Rünger, 2016, "Psychology of Habit". https://dornsife.usc.edu/wendy-wood/wp-content/uploads/sites/183/2023/10/wood.runger.2016.pdf ↩

8 Long Ouyang 외, 2022, "Training language models to follow instructions with human feedback", NeurIPS 2022(OpenAI 공개 글 「Aligning Language Models to Follow Instructions」의 논문). https://arxiv.org/abs/2203.02155 ↩

9 Eric Wallace 외, 2024, "The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions", 프리프린트. https://arxiv.org/abs/2404.13208 ↩

7장 따랐는데도 결과가 어긋나는 이유

1 Jack Clark, Dario Amodei, 2016-12-21, "Faulty Reward Functions in the Wild", OpenAI 블로그. https://web.archive.org/web/20170504013356/https://blog.openai.com/faulty-reward-functions/ ↩

2 Ryan Greenblatt 외, 2024, "Alignment Faking in Large Language Models", arXiv:2412.14093. https://arxiv.org/abs/2412.14093 ↩

9장 어디까지 말하고 어디부터 맡길 것인가

1 H. P. Grice, 1975, “Logic and Conversation”, P. Cole, J. L. Morgan (eds.), *Syntax and Semantics* vol. 3: *Speech Acts*, Academic Press, 41–58. https://projects.illc.uva.nl/inquisitivesemantics/assets/files/papers/Grice1975.pdf ↩

10장 좋은 지시는 어떻게 구성되는가

1 Department of the Army, 2024, *FM 5-0*(계획 수립과 명령 작성 교범, 4 November 2024, Change 1 포함), 5-184항, 6-34항; 확인 브리핑이 받음과 이해를 확인하는 기회라는 설명은 FM 6-0(16 May 2022) 부록 C 8항 Note. https://web.archive.org/web/20251201064154/https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN44590-FM_5-0-001-WEB-3.pdf ↩

2 Department of the Army, 2022, *FM 6-0 Commander and Staff Organization and Operations*(16 May 2022), 부록 C 7–8항. https://web.archive.org/web/20240306081726/https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN35404-FM_6-0-000-WEB-1.pdf ↩

3 Alex B. Haynes 외, 2009, “A Surgical Safety Checklist to Reduce Morbidity and Mortality in a Global Population”, *New England Journal of Medicine* 360(5), 491–499. https://doi.org/10.1056/NEJMsa0810119 ↩

4 Herbert H. Clark, Susan E. Brennan, 1991, “Grounding in Communication”, *Perspectives on Socially Shared Cognition*(American Psychological Association). https://web.stanford.edu/~clark/1990s/Clark,%20H.H.%20_%20Brennan,%20S.E.%20_Grounding%20in%20communication_%201991.pdf ↩

11장 지시의 끝에서 다시 내 의도를 만난다

1 Frederick P. Brooks Jr., 1987, “No Silver Bullet: Essence and Accidents of Software Engineering”, *Computer* 20(4), 10–19. https://archive.org/details/computer-magazine-1987-04 ↩

부록 B 출발 자료와 Intent First의 맥락

1 저자, 2026-10-13, 「내가 물어보면 너는 왜 답하는가 — 대화 원문」. https://dosi.dev/post/2026-10-13-why-do-you-answer-dialogue/ ↩

2 저자, 2026-10-27(2026-10-28 개정), 「내가 지시하면 너는 왜 잘못 행동하는가」. https://dosi.dev/post/2026-10-27-intent-first-agent-instructions/ ↩

참고문헌과 읽는 순서

읽는 순서

관심에 따라 세 경로로 읽을 수 있다. 모두 이 책의 표와 각주에 나온 문헌이다.

- **언어와 의미** — 마르티 외(2023) → 클라크·윌크스깁스(1986) → 클라크·브레넌(1991) → 호프스태터·상데(2013). 같은 낱말이 사람마다 다른 개념을 부른다는 측정에서 시작해, 대화하며 표현을 맞춰 가는 과정을 거쳐, 개념이 유추로 넓어진다는 관점에 닿는다.
- **동기와 수용** — 켈먼(1958) → 밀그램(1963) → 라이언·데시(2000) → 우드·뤼엥거(2016). 요구를 따르는 과정을 나눈 것으로 알려진 구분에서 출발해, 같은 행동을 두고 해석이 갈리는 실험, 외부 요구가 내면화되는 단계, 반복이 만드는 습관으로 이어진다.
- **실무 적용** — 그라이스(1975) → 브룩스(1987) → 미 육군 교범의 백브리프(FM 6-0, 2022) → 헤인스 외(2009). 필요한 만큼 말한다는 대화의 원리에서, 요구는 만들어 써 보기 전에 다 적을 수 없다는 관찰, 받은 쪽이 일을 어떻게 해낼 생각인지 계획 초기에 검토받는 절차, 정해진 시점에 기준을 다시 보게 하는 점검표로 이어진다.

중심 문헌 10선

연도	저자와 문헌	이 글에서의 역할
1958	Herbert C. Kelman, *Compliance, Identification, and Internalization: Three*	켈먼의 구분으로 알려진 순응, 동일시, 내면화의 출처로 인용된다

연도	저자와 문헌	이 글에서의 역할
	Processes of Attitude Change, *Journal of Conflict Resolution* 2(1)	
1986	Herbert H. Clark & Deanna Wilkes-Gibbs, *Referring as a Collaborative Process*, *Cognition* 22(1)	같은 두 사람이 과제를 되풀이하는 동안 대상을 가리키는 낱말이 줄어드는 것을, 함께 받아들인 묘사에 기댄 결과로 설명한다
1987	Michael E. Bratman, *Intention, Plans, and Practical Reason*	의도를 부분 계획과 시간에 걸친 행동 조직의 관점에서 다룬다
1989	Colin Camerer, George Loewenstein & Martin Weber, *The Curse of Knowledge in Economic Settings: An Experimental Analysis*, *Journal of Political Economy* 97(5)	자신이 가진 정보를 떼어 내고 상대를 예상하기 어려운 현상을 분석한다
1991	Herbert H. Clark & Susan E. Brennan, *Grounding in Communication*	대화하는 사람들이 이해의 증거를 주고받으며 '지금 목적에 충분할 만큼' 이해했다고 함께 믿는 데까지 가는 과정을 설명한다
2000	Richard M. Ryan & Edward L. Deci, *Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being*	외부 요구의 내면화와 동기의 여러 형태를 구분한다
2007	Douglas Hofstadter, *I Am a Strange Loop*	자기지시와 '나'라는 상징을 통해 자아를 탐구한다
2008	Steven Pinker, Martin A. Nowak & James J. Lee, *The Logic of Indirect Speech*	간접적 표현의 정보적·관계적 기능과 부인 가능성을 분석한다
2013	Douglas Hofstadter & Emmanuel Sander, *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*	유추와 범주화를 사고의 중심에 놓는다. 한국어판은 『사고의 본질』이다
2023	Louis Marti, Shengyi Wu, Steven T. Piantadosi & Celeste Kidd, *Latent Diversity in Human Concepts*	개인별 개념 차이와 의미 공유에 관한 예상의 차이를 살핀다

본문 사례의 출처

중심 문헌 10선에 없는, 장마다 든 사건·실험·연구 이야기의 1차 출처다(2장 마르티 외 2023, 5장 핑커 외 2008은 위 표에 있다). 각 장의 각주와 같다.

- 1장 — Clayton M. Christensen, Scott D. Anthony, Gerald Berstell, Denise Nitterhouse, 2007, "Finding the Right Job For Your Product", *MIT Sloan Management Review* Spring 2007, "Hiring Milkshakes" 절. https://sloanreview.mit.edu/article/finding-the-right-job-for-your-product/
- 3장 — Mars Climate Orbiter Mishap Investigation Board, 1999, "Mars Climate Orbiter Mishap Investigation Board Phase I Report", NASA, 1999-11-10. https://llis.nasa.gov/llis_lib/pdf/1009464main1_0641-mr.pdf
- 3장 — Mars Climate Orbiter Mishap Investigation Board, 2000, "Report on Project Management in NASA by the Mars Climate Orbiter Mishap Investigation Board", NASA, 2000-03-13, 16쪽. https://discovery.larc.nasa.gov/PDF_FILES/mars_climate_orbiter_phaseII.pdf
- 4장 — National Army Museum, "The order that launched the Charge of the Light Brigade, 1854", NAM. 1962-11-4-3. https://collection.nam.ac.uk/detail.php?acc=1962-11-4-3
- 4장 — Hansard, HL Deb 2 March 1855 vol 137 cc3-7, "The Earl of Lucan"(루칸의 1854-11-30 편지 낭독). https://api.parliament.uk/historic-hansard/lords/1855/mar/02/the-earl-of-lucan

- 6장 — Stanley Milgram, 1963, “Behavioral Study of Obedience”, *Journal of Abnormal and Social Psychology* 67(4), 371–378. https://www.psy.miami.edu/_assets/pdf/rpo-articles/milgram-1963.pdf
- 6장 — Stanley Milgram, 1965, “Some Conditions of Obedience and Disobedience to Authority”, *Human Relations* 18(1), 57–76. https://web.mit.edu/curhan/www/docs/Articles/15341_Readings/Influence_Compliance/Milgram_Some_Conditions_of_Obedience_and_Disobedience_to_Authority.pdf
- 6장 — S. Alexander Haslam, Stephen D. Reicher, Megan E. Birney, 2014, “Nothing by Mere Authority”, *Journal of Social Issues* 70(3), 473–488. http://hdl.handle.net/10034/604991
- 7장 — Jack Clark, Dario Amodei, 2016-12-21, “Faulty Reward Functions in the Wild”, OpenAI 블로그. https://web.archive.org/web/20170504013356/https://blog.openai.com/faulty-reward-functions/
- 7장 — Ryan Greenblatt 외, 2024, “Alignment Faking in Large Language Models”, arXiv:2412.14093. https://arxiv.org/abs/2412.14093
- 9장 — H. P. Grice, 1975, “Logic and Conversation”, P. Cole, J. L. Morgan (eds.), *Syntax and Semantics* vol. 3: *Speech Acts*, Academic Press, 41–58. https://projects.illc.uva.nl/inquisitivesemantics/assets/files/papers/Grice1975.pdf

- 10장 — Department of the Army, 2022, *FM 6-0 Commander and Staff Organization and Operations*(16 May 2022), 부록 C 7–8항. https://web.archive.org/web/20240306081726/https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN35404-FM_6-0-000-WEB-1.pdf
- 10장 — Department of the Army, 2024, *FM 5-0*(계획 수립과 명령 작성 교범, 4 November 2024, Change 1 포함), 5-184항, 6-34항. https://web.archive.org/web/20251201064154/https://armypubs.army.mil/epubs/DR_pubs/DR_a/ARN44590-FM_5-0-001-WEB-3.pdf
- 10장 — Alex B. Haynes 외, 2009, "A Surgical Safety Checklist to Reduce Morbidity and Mortality in a Global Population", *New England Journal of Medicine* 360(5), 491–499. https://doi.org/10.1056/NEJMsa0810119
- 11장 — Frederick P. Brooks Jr., 1987, "No Silver Bullet: Essence and Accidents of Software Engineering", *Computer* 20(4), 10–19. https://archive.org/details/computer-magazine-1987-04

함께 참조한 자료

- Douglas Hofstadter, MIT 강연 *Analogy as the Core of Cognition* (2005). 이름이 있는 개념과 이름이 없는 개념, 여러 추상화 수준에서의 유추를 설명한다.
- Bertram H. Raven, Joseph Schwarzwald & Meni Koslowsky, *Conceptualizing and Measuring a Power/Interaction Model of Interpersonal Influence* (1998). 영향력의 여러 기반을 구분한다.

- Wendy Wood & Dennis Rünger, *Psychology of Habit* (2016). 맥락에 따라 학습한 습관과 목표 지향적 행동의 관계를 다룬다.

- Long Ouyang 외, *Training Language Models to Follow Instructions with Human Feedback* (2022, NeurIPS 2022). 사람이 쓴 시범 답과, 사람이 매긴 출력 순위로 만든 보상 모형으로 지시 수행을 학습시킨 공개 사례다.

- Eric Wallace 외, *The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions* (2024, 프리프린트). 지시의 출처에 따라 우선순위를 학습시키고, 위 단계와 맞는 아래 단계 지시는 따르되 어긋나거나 상관없는 지시는 무시하게 하자고 제안한다.

- dosi.dev, 「내가 물어보면 너는 왜 답하는가 — 대화 원문」(2026-10-13). 이 글의 수용·응답·판단 권한에 관한 출발점이다.

- dosi.dev, 「내가 지시하면 너는 왜 잘못 행동하는가」(2026-10-27 공개, 2026-10-28 개정). 지시와 의도의 간극, 네 원인과 처방, 원문과 추론을 구분하는 실무 논의의 출발점이다.

한국어판 서지 확인: KAIST 도서관 — 『사고의 본질』.

- Merriam-Webster, "agent." 행위자·대리인·컴퓨터 응용 프로그램의 용례와 라틴어 어원을 확인한 사전 항목.

- The American Heritage Dictionary, "agent." 라틴어 동사 *agere*, 현재분사 *agens*, 어간 *agent-* 사이의 관계와 행위자·대리인의 의미를 확인한 사전 항목.