AI 비서 팀에게 일을 맡기다 겪은 사고 9건을 '사람이 승인할 일 vs AI가 혼자 할 일' 한 장짜리 규칙표로 바꾼 이야기

분야: 업무 생산성 개선을 위한 AI 활용
활용한 AI 도구: OpenClaw(오픈소스 개인 AI 비서 실행 도구: 메인 비서 1 + 전담 에이전트 + 정해진 시간 작업), Anthropic Claude(Opus — 대화·판단·최종 문장 / Sonnet — 반복 작업 / 최상위 모델 — 틀리면 안 되는 최종 확인), 기억용 Markdown 파일(일지·장기 기억·규칙 파일), SHA-256 해시 확인 스크립트(PowerShell/bash)
한 줄 소개: 본업이 있는 직장인이 AI 비서 여러 개를 약 한 달 매일 굴리며 겪은 실패를 '증상→원인→재발 방지 규칙'으로 기록하고, 누구나 복사해 쓸 수 있는 안전 규칙표와 복구 순서로 정리했습니다.

■ 한 줄 요약
어느 날 AI 비서가 화면 좌표로 짐작해 누른 메뉴가 '로그아웃'이었고, 계정 세 개가 한꺼번에 로그아웃됐습니다. AI가 모자라서가 아니었습니다. 제가 "여기서부터는 멈추고 물어봐"라는 선을 그어 두지 않았기 때문입니다. 그 뒤로 사고가 날 때마다 규칙을 한 줄씩 적었고, 그 규칙을 '사람이 승인할 일 vs AI가 혼자 할 일' 한 장짜리 표로 만들었습니다.

---

① 어떤 상황에서 AI를 활용했나요?
저는 평일에 본업이 있는 직장인입니다. 퇴근 뒤와 주말에 하는 일(자료 조사, 보고서·글 초안, 파일 정리, 정기 점검, 웹 관리 화면 작업)을 AI 비서에게 맡겼습니다.
처음 부딪힌 문제는 두 가지였습니다.
· 비서가 어제를 모릅니다. 어제 "확인도 안 하고 됐다고 하지 마"라고 고쳐 줬는데, 다음 날 대화가 새로 시작되자 비서는 또 확인 없이 "됐습니다"라고 했습니다.
· 맡길수록 사고가 납니다. 시킨 일을 끝내려고 애쓰다가 사람이라면 멈췄을 지점을 그냥 지나갑니다. 두 에이전트가 같은 브라우저를 동시에 쓰다가, 한쪽이 바꾼 계정으로 다른 쪽이 일을 이어 간 적도 있습니다.
그래서 제 질문은 "AI로 무엇을 더 빨리 할까"가 아니라 "AI에게 어디까지 맡겨도 안전한가"였습니다.

---

② 어떤 AI를 어떻게 활용했나요?
【구성】 메인 비서 1(저와 대화·판단) + 전담 에이전트 7(업로드·정리·조사 같은 반복 작업) + 정해진 시간 작업 19개(백업·점검·알림). 메인이 일을 나눠 맡기고 결과를 받아 옵니다. 도구는 오픈소스 AI 비서 실행 도구 OpenClaw와 Anthropic Claude 모델입니다.
【이 도구를 고른 이유】 비서의 기억과 규칙이 모두 내 컴퓨터의 일반 텍스트 파일이라, 비서가 무엇을 기억하고 어떤 규칙을 따르는지 사람이 직접 열어 보고 고칠 수 있습니다. 대화창형 AI는 편하지만 '어제 고쳐 준 것'을 내가 확인할 방법이 마땅치 않았습니다.

【모델을 일에 맞게 나눔】 처음에는 모든 일을 가장 비싼 모델 하나로 돌리다가 일주일 사용 한도의 대부분을 써 버렸습니다. 지금은 대화·판단과 사람에게 보이는 문장은 상위 모델, 반복 작업은 한 단계 낮은 모델, 틀리면 안 되는 최종 확인만 최상위 모델이 맡습니다. 이 배분은 규칙 파일에 표로 적어 두었습니다.
【기억은 파일 세 층】 일지(그날 한 일) / 장기 기억(오래 갈 결정과 그 이유) / 규칙 파일(규칙과 바뀌면 안 되는 약속). 장기 기억에는 "다음에 다르게 행동해야 하는 것"만 올리고, 비밀번호·인증 코드·다른 사람의 개인정보·짐작은 어디에도 적지 않습니다.
【사고 → 규칙 고리】 사고가 나면 "증상 → 원인 → 고친 것 → 재발 방지" 네 줄로 기록하고, 재발 방지 줄을 규칙 파일에 넣습니다.

【실제 업무 장면 — 조사 보고서 초안을 AI가 쓰고 사람이 확인】
이 공모전에 내기까지의 과정도 같은 방식이었습니다.
1. 비서가 공모전 목록 공고 430건을 훑고, 조건(마감 전·개인 참가·현금 상금·AI 사용 가능)에 맞을 만한 약 35건은 상세 공고를 직접 열어 확인한 뒤, 우선 후보 5건과 그 밖의 후보 10건으로 정리했습니다.
2. 그중 8건은 주최 측 원문 자료(공고문 PDF·포스터·공식 게시글)를 직접 받아, 마감·제출물·배점·유의사항을 원문 그대로 인용한 조사 보고서 초안을 썼습니다.
3. 초안은 출처끼리 다른 곳을 따로 표시했습니다. 예를 들어 이 공모전 제출 사이트 첫 화면에는 지난 시즌의 마감일과 다른 심사 기준이 남아 있었는데, 비서는 이를 '예전 공지로 보임'이라고 표시하고 공식 안내 PDF의 배점을 따르자고 제안했습니다. 확인하지 못한 것(정확한 마감 시각 등)도 따로 적었습니다.
4. 저는 보고서를 읽고 낼 곳과 따를 기준만 정했습니다. 원문을 하나하나 찾아 읽는 일은 비서가, 판단은 제가 했습니다.

---

③ 활용 결과 어떤 변화가 있었나요?
【실제로 겪은 사고 9건과 그 뒤 생긴 규칙】
1. 화면 좌표로 짐작해 누른 메뉴가 '로그아웃'이라 계정이 전부 로그아웃됨 → 계정·설정 메뉴는 버튼 이름을 확인한 뒤에만 누른다. '로그아웃'은 비서가 누르지 않는다.
2. 두 에이전트가 같은 브라우저를 동시에 써서 활성 계정이 서로 바뀜 → 같은 브라우저·계정을 쓰는 작업은 한 번에 하나만. 중요한 작업 직전에 계정 이름을 다시 확인한다.
3. 하위 에이전트의 "완료했습니다"를 그대로 전했는데 결과가 틀림 → 증거(파일 경로·명령 출력·해시·화면 캡처)를 함께 받고, 사람에게 건네기 전에 메인이 직접 열거나 다시 계산한다.
4. 응답 수가 대상 문항 수와 맞지 않는데 결과를 엶 → 결과를 열기 전에 개수부터 센다. ID·번호는 추측하지 않고 출처에서 복사한다.
5. 올리면서 바로 공개 예약했다가 저작권 문제가 붙음 → 비공개로 먼저 올리고, 검사가 "문제 없

음"으로 끝난 걸 확인한 뒤에만 공개한다.
6. 비서가 자기가 돌아가는 프로그램을 업데이트하다 같이 끊김 → 비서는 자기 자신을 끄거나 재시작·업데이트하지 않는다. 안내까지만 한다.
7. 브라우저 도구 고장이 재시작으로 안 고쳐짐(남은 프로세스가 원인) → 사람에게 재시작부터 부탁하지 말고 원인을 먼저 찾는다. 안전한 조치는 비서가 직접 하되, 보안 허용 창을 대신 누르지는 않는다.
8. 일주일 사용 한도가 거의 바닥남 → ②의 모델 배분 규칙이 생김.
9. 결과물이 시스템 드라이브를 가득 채움 → 결과물·임시 파일을 둘 '한 곳'을 정해 규칙 파일에 적음.

【숫자로 본 한 달】 (모두 운영 기록에서 직접 센 값입니다)
· 운영 기간: 일지 파일 30개(9월 30일~10월 29일)
· 일을 나눠 맡긴 횟수: 하위 작업 약 420건
· 사람이 바로잡아 준 것을 규칙으로 바꾼 것: 21건(비서의 '고쳐 받은 것' 목록)
· 규칙 파일 22개 절·54줄, 장기 기억 파일 53개

【바뀐 점】
· 사고가 같은 형식으로 쌓이니 공통 원인이 보였습니다. 대부분 "확인 없이 짐작한 것"(좌표, 보고, 개수, 링크)이었습니다.
· 위험한 일은 비서가 "준비까지만" 하고 저에게 묻습니다. 다른 공모전에 음원을 낼 때도 하위 에이전트가 신청서 칸을 모두 채우고 '제출' 버튼 바로 앞에서 멈췄습니다. 메인 비서가 칸마다 내용을 다시 확인했고, 제출은 제가 곡을 듣고 OK한 뒤에만 했습니다. 저는 결과물을 처음부터 만드는 사람에서, 준비된 것을 확인하고 승인하는 사람으로 바뀌었습니다.
· 솔직히 말씀드리면, 몇 분을 아꼈는지 같은 시간 측정은 하지 않았습니다. 그래서 시간 절약 숫자는 적지 않았습니다.

---

④ 나만의 방식 또는 개선 포인트는 무엇인가요?
【1. 한 장짜리 규칙표 — 사람이 승인할 일 vs AI가 혼자 할 일】 (첨부 이미지)
· AI 혼자 해도 됨: 파일 읽기·정리, 검색, 초안 쓰기, 기억 파일 정리, 상태 점검 → 끝까지 하고 결과 보고
· 조건부: 비서가 띄운 보조 프로그램 재시작, 임시 파일 정리 → 원인 확인 후 진행, 한 일을 기록
· 준비 후 승인: 게시·공개·예약, 대량 발송, 삭제·덮어쓰기, 결제·구독 → 비공개로 준비하고 "이번 건 OK?"를 물음. 승인은 그 한 건에만 유효
· 사람만: 보안·권한 확인 창, 로그인·로그아웃, 계정 전환, 비밀번호·인증 코드 입력 → 비서는 멈추고 보고

· 하지 않음: 남의 개인정보를 공개 채널에 올리기, 의료·법률·투자 결정
【2. 바뀌면 안 되는 약속은 '헌장'으로 떼어 내고 지문(해시)으로 확인】
비서의 성격·말투 파일은 매일 고쳐지다 보니 중요한 약속까지 흐려질 수 있었습니다. 그래서 절대 바뀌면 안 되는 몇 줄(예: "사실을 지어내지 않는다")을 따로 떼고 SHA-256 지문을 새겨 두었습니다. 비서는 일을 시작할 때 지문을 비교하고, 다르면 아무것도 고치지 않고 사람에게 알립니다. 한계도 적어 두었습니다. 지문은 "바뀌었다"만 알려 줄 뿐, 비서가 약속을 지키는지는 알려 주지 않습니다.
【3. 어떤 사고든 같은 복구 순서】 멈춤 → 재현 → 원인 좁히기(최근에 바뀐 것부터) → 최소한만 고치기 → 같은 방법으로 다시 확인 → 재발 방지 한 줄 기록. "고쳤다"는 말은 다시 확인한 뒤에만 합니다.
【4. 밤 시간 알림 규칙】 23:00~08:00에는 급한 일만 보내고, 새로운 게 없으면 보내지 않습니다. 알림이 잦으면 사람은 결국 알림을 끄고 진짜 중요한 알림을 놓치기 때문입니다.

---

⑤ 다른 사람도 따라 할 수 있나요?
네. 아래 9줄을 AI 비서의 규칙 파일 맨 위에 붙여 넣고 상황에 맞게 고치면 됩니다. 다른 AI 도구라도 '맞춤 지시' 칸이 있다면 옮겨 쓸 수 있습니다.
【복사해서 쓰는 안전 규칙】
1. 비밀번호·API 키·인증 코드는 기억 파일·일지·지시문에 적지 않는다.
2. 보안·권한 확인 창은 누르지 않는다. 멈추고 보고한다. 허락은 그 한 건에만 유효하다.
3. 공개·삭제·돈·계정 변경은 준비까지만 한다. 승인 후에만 실행한다. 삭제는 휴지통으로.
4. 계정·설정 메뉴는 화면 좌표로 누르지 않는다. 버튼 이름을 확인한 뒤에만 누른다. '로그아웃'은 누르지 않는다.
5. "됐다"고 말하기 전에 직접 확인한다. 확인 못 했으면 못 했다고 말한다.
6. 하위 에이전트에게는 증거(경로·출력·해시)를 함께 받는다. 중요한 건 다시 확인한다.
7. 23:00~08:00에는 급한 일만 보낸다. 새로운 게 없으면 보내지 않는다.
8. 도구가 고장 나면 우회하지 말고 원인을 먼저 찾는다. 안전한 조치는 직접 하고, 보안 창·사람의 개인 브라우저·되돌리기 어려운 일은 보고한다.
9. 모르는 것, 최신 정보는 기억으로 답하지 않고 먼저 찾아본다.
【시작하는 3단계】
1) 위 9줄을 붙여 넣는다.
2) 사고가 나면 "증상 → 원인 → 고친 것 → 재발 방지" 네 줄로 적는다.
3) 재발 방지 줄을 규칙 파일에 넣고, 같은 주제의 옛 규칙은 고쳐 쓴다(한 주제에 규칙은 한 줄만).

■ 윤리·책임 — 이 사례에서 지킨 것
· 개인정보: 이 글에는 이름·계정·파일 경로 같은 식별 정보를 넣지 않았습니다. 비서의 기억 파일에

도 비밀번호와 다른 사람의 개인정보를 적지 않습니다.
· 사람의 최종 결정: 공개·삭제·결제·계정 변경은 사람이 승인하고, 보안 확인 창은 사람만 누릅니다.
· 정확성: "됐다"는 말과 하위 에이전트 보고는 증거로 확인합니다. 확인하지 못한 것은 확인하지 못했다고 씁니다.
· 저작권: 공개 전에 플랫폼 저작권 검사를 통과한 뒤에만 공개합니다.
· 한계: 한 사람이 약 한 달 운영한 경험이라 일반화에는 한계가 있고, 시간 절약 효과는 측정하지 않았습니다.
· AI 사용 표기: 이 글은 제 운영 기록을 바탕으로 AI 비서(Claude)가 초안을 쓰고, 제가 사실을 확인하고 고쳤습니다.

■ 참고 자료(출처)
· OpenClaw 공식 문서 — 메모리: https://docs.openclaw.ai/ko/concepts/memory
· OpenClaw 공식 문서 — 보안(개인 비서 1인 신뢰 모델): https://docs.openclaw.ai/ko/gateway/security
· OpenClaw 공식 문서 — 실행 승인: https://docs.openclaw.ai/ko/tools/exec-approvals
· OpenClaw 공식 문서 — 에이전트 작업 공간·git 백업: https://docs.openclaw.ai/ko/concepts/agent-workspace
(첨부: 한 장짜리 규칙표 이미지, 본문 PDF)