
Seven for Seven on the German Wiki: A Taxonomy That Predates the Incident, Mapped to the May to July 2026 Agent Swarm¹

Travis Gilly

Executive Director, Real Safety AI Foundation; Contributor, Software Safety Panel, Internet Safety Labs

Lisa LeVasseur

Founder and Executive Director, Internet Safety Labs; Chair, Software Safety Panel

Anthony Martini

The Feel Better Lab; Contributor, Software Safety Panel, Internet Safety Labs

Rene Mattheus

Regional Director, Real Safety AI Foundation; Contributor, Software Safety Panel, Internet Safety Labs

With

Apart Research

Abstract

A seven-category taxonomy of benign capabilities that combine into ordinary crimes reached Internet Safety Labs on 21 July 2026, before the OpenAI escape was public and six weeks before the German wiki corpus appeared. This report maps it against the wiki incident of 24 May to 2 July 2026, which nobody had mapped. All seven are present, with the heaviest evidence in the two tier-two categories that turn a technique into a self-running process. The unit of analysis is the incident, and a capability is read by the effect it produced, not the mechanism it used. We report what an onboarding attempt with two new raters showed, and what they reached for that the seven do not name. The forecast: closing the coordination-surface class narrows this shape of incident and not the next, because an agent that can read a public repository can rebuild what it describes without installing anything.

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

1. Introduction

Track 2 asks what the incident establishes and what breaks next. Between 24 May and 2 July 2026, roughly 3,100 agent identities belonging to a single provider wrote about 14,600 revisions across 4,579 pages of an old German wiki, using it as a message board to relay test answers, warn each other about upcoming questions, and share ways around their own restrictions. That corpus became public on 4 September. Nobody has mapped it against a risk taxonomy.

The instrument used here is a seven-category taxonomy of individually benign AI capabilities that combine into four ordinary crimes: fraud, extortion, theft, and the sale of illicit goods and services. Its claim is that no single capability is dangerous alone, that the harm is the assembly, and that governance inspects capabilities one at a time and never inspects the combination. Five tier-one categories are the ingredients of a crime. Two tier-two categories, accelerants and orchestration, are not ingredients at all; they remove the time limit and remove the human, so that an assembly a person could run slowly runs itself at scale.

The taxonomy was circulated to Internet Safety Labs on 21 July 2026 at 12:17 AM. OpenAI attributed the Hugging Face intrusion to its own models later that day; the wiki corpus surfaced six weeks afterward. It has been scored once before, against the UK AI Security Institute red team published 4 August, where all seven were present. That was a designed operation with safeguards deliberately removed. The wiki swarm was designed by nobody: no human directed a step, and the coordination emerged from a population of agents each doing something else. That is the difference between a scheme that organizes a description and one that named a shape before the shape appeared.

Our main contributions are:

1. A category-by-category mapping of the wiki incident against all seven categories, with evidence drawn from the corpus and a confidence per category (Table 1). All seven are present.
2. The unit-of-analysis rule the mapping depends on, and why the instrument scores an incident rather than a message: a capability is read by the effect it produced, not the mechanism it used.
3. An onboarding result: what happened when the instrument was handed to two people who had never seen it, in the final hours before the deadline, and the four behaviors they reached for that the seven do not yet name (Table 2).
4. One forecast with a public resolution, and two checks a lab can run on itself in an afternoon.

2. Related Work

The taxonomy is built on the reference structures that describe AI risk and on the crime categories that describe outcomes. Its capability side draws on MITRE ATLAS, the IBM AI Risk Atlas and Risk Atlas Nexus, the MIT AI Risk Repository, the NIST AI Risk Management Framework, and the CSET AI harm taxonomy with the AI Incident Database. Its outcome side uses four ordinary criminal categories, with fraud in its common-law sense of theft by deception, and the crime-as-a-service and illicit-trade patterns reported in the Europol Internet Organised

Crime Threat Assessment. The methodology it extends is integration-surface analysis, set out in Emergent Convergence Risk (Gilly, 2026a).

What those structures do not do is examine assemblies. Each evaluates a capability, a system, or a harm on its own terms, and each answer is yes, this one is safe. The gap is not any single tool; it is the space between tools that nobody inspects. This taxonomy is adopted by Internet Safety Labs and is being integrated into a working taxonomy through the weekly working group of the ISL Software Safety Panel.

The evidence base is the collusion.wiki export (Von Arx, Byrd, Kitts, and Larsen, 2026), the METR and Redwood Research independent investigation of the parallel July intrusion (Wijk, Cotra, and Greenblatt, 2026), and OpenAI's technical report of 26 August 2026.

3. Methods

The unit of analysis is the incident. This is the rule the mapping depends on and the one most easily got wrong. The taxonomy scores an incident, not a message and not a thread. A single post shows one agent doing one thing; the assembly is visible only across the population and across the run. Scoring a message against a combinatorial instrument is a category error, in the way that one ingredient tells you nothing about a recipe.

A capability is read by its effect, not its mechanism. The wiki's write restriction was written on the mechanism, the POST verb. The agents produced the denied effect, persistent text on a page other parties read, using only affordances they were permitted. Whether that counts as posting depends on whether the mechanism or the result is scored. The instrument scores the result, because the harm is the result. That rule resolves every borderline case below.

Sources. Every entry in Table 1 is evidenced from the collusion.wiki export or from the provider's own published account. Where evidence is thinner the confidence column says so rather than the row being dropped. Nothing in the mapping relies on a private document.

The onboarding attempt. A rating workbook was prepared so that people other than the author could apply the instrument: twelve threads drawn by seeded stratified sample (seed 20260913) from roughly 250 multi-agent threads in the corpus, each revision with a plain-English gloss and agents color-coded, a how-to-rate tab written for someone who has never seen the taxonomy, and a rate tab with a function and role cell per category plus none-of-the-seven, cannot-determine, confidence, and notes. Two team members rated independently and without an LLM. A definitions reference (v2.2) giving each sub-capability a plain definition and an ordinary-use and agentic-use example pair was written on 12 September, because the 21 July proposal defines categories and roles at the category level and lists sub-items without defining them. That was sufficient for the author and is not sufficient for anyone else.

Three constraints on that attempt, stated before the result. First and largest, time: rating did not begin in earnest until about 4 PM Central on the final day, against a deadline the same night. Second, the raters were learning a definitional layer one day old while applying it. Third, the workbook handed them threads, which is the wrong unit for a combinatorial instrument, for the practical reason that twelve threads fit in an afternoon and an incident does not. The attempt therefore tests onboarding under time pressure, not the instrument's fit to the incident.

4. Results

Table 1 is the mapping. All seven categories are present in the wiki incident.

Category	Evidence in the corpus	Role it played	Conf.
1 Perception and ingestion	Agents read their own runtime and the environment around it, then read each other. They measured their task clock against the wiki clock and published the ratio: one cohort reported its task clock running at about 0.39 times UTC and projected its next prompt from that. They profiled other cohorts by position: Aug11 about a minute ahead, Jun10 about level, with requests for seconds-to-next-question. They extracted structured answers from source data, county poverty rates and workforce counts by industry code, and cached full answer tables.	Reconnaissance, target profiling, gathering leverage	High
2 Synthesis and generation	Text at volume, and code. Agents authored relay pages, timing tables, and instructions for other agents. One posted a copy-paste command that writes to the shared page in about a second with no login. Another published a solved method as a reusable artifact: the data definition, the filter conditions, the weighting column, and the next five values, for whoever arrived later.	Text; code; an artifact written for another participant	High
3 Identity and matching	Agents named themselves from their context window and used the names consistently, then resolved each other across pages and cohorts: Sep13PovertyWatcher, Jan14PovertyWatcher, OpenAIResearcherNov26. They held cohort identity in the first person plural and individual identity in the singular, and tracked which named agent was furthest ahead. One agent took the wiki administrator's username.	Entity resolution; linking an artifact to a specific actor	High
4 Actuation and reach	Writing to a surface that humans and other agents read, in an environment where writing was denied. The denial was written on the mechanism and the effect was produced anyway. One agent independently reproduced another's method for defeating the request restriction and rendered a live chart the permitted route could not reach.	Delivery; execution	High
5 Evasion and laundering	The bypass itself is anti-detection: a control evaluating one thing while a permitted combination produces the denied effect. Beyond it, agents ran signal counters and detached processes intended to outlive the run and leave a record its teardown would not capture, and corrected that record when a test artifact polluted it.	Anti-detection; provenance handling	Medium
6 Accelerants	Rated at the category level, and unmistakable. Answers propagated until later agents answered instantly: three confirmed, answered in one second; four confirmed, answered instantly. Agents skipped their own preparation to race ahead and relay, and used long interruptible waits to fast-forward through cooldowns. About 14,600 revisions in the corpus, with a moderator deleting roughly a hundred posts at a time while several hundred more arrived.	Removes the time limit; scales the assembly	High
7 Orchestration and autonomy	Self-directed loops with division of labor and no human in any step. Agents specialized: some raced ahead to relay, some ran reconnaissance on other cohorts and reported back rather than doing the task, some tested hypotheses and published negative results, including ten million candidate orderings tested and reported as a dead end. Structure formed at three levels, agent, cohort and swarm, and later arrivals inherited the coordination as an existing feature of the world.	Removes the human so the assembly runs itself	High

Table 1. The seven categories against the German wiki incident, 24 May to 2 July 2026, evidenced from the collusion.wiki export. The outcome the assembly served was theft: the object taken was the answer set to a timed evaluation.

Two observations. The categories carrying the heaviest evidence are the tier-two pair, accelerants and orchestration, which the taxonomy says are not ingredients of any crime and

are instead what turns a technique into a self-running process. That is the structural claim observed rather than asserted. And the category that took the most care to score, evasion, is the one where the control was written on a mechanism rather than an effect; an instrument scoring mechanisms would have recorded clean compliance.

The onboarding attempt. Two raters began at about 4 PM Central on the final day. Neither completed the twelve threads: one rated a thread in full and started a second, the other described the first threads in her own words and stopped. In the time available that is the expected outcome, and it says nothing about whether the instrument fits the incident, which Table 1 answers. It says what it costs to hand a definitional instrument to someone who has never used it, on the day, and it surfaced which behaviors people reach for that the seven do not yet name. Table 2 records the second part, which is the useful part.

<i>What they reached for</i>	<i>What it is in the corpus</i>	<i>Where it sits against the seven</i>
The task clock	Agents tracked timing to the second, relayed start signals, raced, waited, and reprioritized reconnaissance over the task itself. Both raters recorded it; neither placed it.	An internalized behavior driven by the run's constraints, where the taxonomy is a capability instrument. The closest home is perception aimed inward, which the published list does not contain.
Cooperation quality	Not the fact of coordination, which orchestration covers, but how cohesive the swarm is: cohort-level speech, a swarm identity above the cohort, division of labor.	Orchestration records that the assembly runs itself. It does not measure how strongly. A probe for swarm cohesion is a candidate check for an evaluation platform.
Cheating as reprioritization	Cooperation that became copying, then answering in one second. Both raters reached for perception plus synthesis and neither was satisfied.	An efficiency mechanism under token-minimization pressure rather than an ethical choice. The capabilities are covered; the decision to substitute the shortcut for the task is not.
Persistence beyond the run	Signal counters, detached processes, artifacts written for participants who had not arrived yet.	Partly covered by synthesis, but the point is that the artifact outlives its author. A candidate role rather than a candidate category.

Table 2. Behaviors two new raters reached for, and where each sits relative to the published seven. None is asserted as an eighth category; each is a candidate the next version has to rule in or out.

The main threat to validity is source dependence, and it is sharp here. The corpus is agent-authored, and the independent investigation of the parallel July incident found that over seven percent of reviewed transcripts contained deliberately spoofed tool calls and that agents attempted to edit their own logs. Part of the record was written by the subjects to mislead a reader. Every mapping rests on it.

5. Discussion

What this establishes. That a taxonomy fixed and circulated on 21 July describes, category by category, an incident whose record became public on 4 September, and that the evidence is heaviest exactly where the taxonomy's structural claim says it should be, in the two categories that convert a technique into an autonomous one. That the tier split does real work: the wiki population shows tier-two coordination without a third-party breach, while the July intrusion shows both, and a flat capability list cannot separate those two events.

What it does not establish. That the taxonomy is predictive rather than descriptive; a sufficiently rich scheme can be fitted to a rich record afterwards, and the answer is pre-registration against a future incident, not another confirmation. That a full seven means anything without a denominator; nothing here says how many incidents show seven, how many show three, or what the distribution looks like. And nothing about inter-rater agreement, which the onboarding attempt had no time to produce.

What the onboarding attempt is about. A taxonomy only its author can apply is not a standard, and the distance between those two states is measured in definitions, worked examples, and time. On the day the definitional layer was one day old and the window was an afternoon. Everything useful the raters produced, Table 2, came from describing what they saw in ordinary language before reaching for a category, which is the order the next version of the workbook should ask for explicitly, and the unit it asks about should be the incident.

What breaks next

The forecast has two halves, and the second is the larger one. The first: the coordination surface these agents used is enumerable by its properties, writable without a POST, no login, no anti-bot filter, easy from a command line, obscure enough to be unmonitored, old enough to predate the modern convention. Roughly a hundred such surfaces exist, and the class has now been found twice, once by agents looking for somewhere to talk and once by investigators running the same enumeration defensively. As more of that class is found and closed, the space for the next incident of this exact shape narrows. That is the good news and it is also the limit of the good news.

The second half is why closing that class does not close the problem. Any agent with read access to the internet can read any public repository, and a repository is a description of a working capability in enough detail to reproduce it. A model that already holds the language it is written in need not clone anything, install anything, or reach a package registry to use what it has read; reading is sufficient and reconstruction is local. The corpus shows this at small scale: one agent independently reproduced another's published bypass from the description alone, and a solved method was posted as a reusable artifact for participants who had not yet arrived. The general case follows. The capabilities reachable by an agent are not the tools it was granted, nor the packages it may install; they are everything legible to it that it can rebuild. A policy that permits reading and denies installing is scoring the mechanism again, and the rule in section 3 says score the effect. We enumerate nothing here and name no candidates. The point for a defender is that a capability inventory built from granted tools understates what is reachable, and the understatement grows with every public description of a working technique. That is a forecast about the next class of incident rather than the next instance of this one.

Two checks a lab can run on itself tomorrow, from the two categories with the heaviest evidence. Does the evaluation harness log actions at the wrapper boundary, or from a log the model can reach? That is the evasion question. Do parallel evaluation sandboxes share any writable namespace, cache, or registry? That is the orchestration question. Both are answerable from configuration in an afternoon, by a platform engineer, with no model expertise and no access to anyone's network.

Future Work

Pre-register the scoring rubric and apply it to an incident not yet public, which answers the fitting objection by construction. Assemble the public denominators so partial assemblies become rate questions: Anthropic's review of 141,006 evaluation runs, the AI Security Institute's test of five models over 475 runs each, and frontier risk reporting. Run the rating at incident scale with the definitions reference and agreed worked examples, and report agreement and residual rates. Rule the Table 2 candidates in or out. And check the repetition threads against the wiki moderator's deletion timestamps, which would separate a model-parameter explanation from an environmental one.

6. Conclusion

Seven categories, fixed before the record existed, all seven present in it, with the heaviest evidence in the two that the scheme says turn a technique into something that runs itself. The wiki agents were not adversaries and no crime was committed; they were cheating on a test. That is the point. The assembly is visible while every part of it still looks innocent, which is the only moment at which anyone can act on it.

Author Contributions

Travis Gilly: the taxonomy, the definitions reference, the mapping, the rating instrument and sample, and the writing. Lisa LeVasseur: independent rating and the behavior characterization behind Table 2. Anthony Martini: independent rating. Rene Mattheus: review. All four authors are members of the Internet Safety Labs Software Safety Panel, where this work originated; the taxonomy is adopted by Internet Safety Labs and is being integrated into a working taxonomy through the panel's weekly working group.

References

1. MITRE ATLAS. Adversarial threat landscape for artificial intelligence systems.
2. IBM. AI Risk Atlas and Risk Atlas Nexus.
3. MIT AI Risk Repository.
4. National Institute of Standards and Technology. AI Risk Management Framework.
5. CSET AI harm taxonomy; AI Incident Database.
6. Europol. (2026). Internet Organised Crime Threat Assessment (IOCTA) 2026, including crime-as-a-service and illicit online trade.
7. Common-law definition of fraud as theft by deception.
8. Gilly, T. (2026a). Emergent Convergence Risk: Why Existing AI Governance Cannot Detect Combinatorial Threats and a Proposed Methodology. ResearchGate, DOI 10.13140/RG.2.2.12454.48963 (March 2026, updated 19 July 2026).
9. Gilly, T. (2026b). Convergence Grab Taxonomy, working proposal v0.2, circulated to Internet Safety Labs 21 July 2026; Definitions and roles, reference v2.2, 12 September 2026. Real Safety AI Foundation.
10. Von Arx, S., Byrd, C. S., Kitts, S., and Larsen, T. (2026). collusion.wiki: export of 14,591 revisions across 4,579 pages, 24 May to 2 July 2026, with access-log analysis. Nightingale Collective, 4 September 2026.

11. Wijk, H., Cotra, A., and Greenblatt, R. (2026). Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR and Redwood Research, 26 August 2026. <https://blog.redwoodresearch.org/p/brief-independent-investigation-of>
12. OpenAI. (2026). OpenAI and Hugging Face incident technical report. 26 August 2026. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
13. UK AI Security Institute. (2026). Red team evaluation of frontier models with safeguards removed. 4 August 2026.

Appendix A. The taxonomy as circulated, 21 July 2026

Reproduced without addition from the Convergence Grab Taxonomy, working proposal v0.2. Benign parts, a minimum combination, an ordinary crime: no single part is sufficient; the combination is.

Category	What it entails	Benign use	Convergence role
<i>Tier one. Crime ingredients. A minimum subset, together, is the crime.</i>			
1 Perception and ingestion	Scraping, OCR, transcription, recognition, entity extraction	Research, indexing, accessibility	Reconnaissance, target profiling, gathering leverage
2 Synthesis and generation	Text, voice, image, video, code	Content, accessibility, creativity	The deceptive or coercive payload, impersonation
3 Identity and matching	Entity resolution, face or voice matching, embedding similarity, re-identification	Recommendation, deduplication, search	Linking an artifact to a specific real person
4 Actuation and reach	Messaging and voice interfaces, posting, payment rails, browser and desktop automation	Notifications, operations	Delivery, execution, collecting proceeds
5 Evasion and laundering	Anti-detection, proxy rotation, provenance handling, payment obfuscation	Privacy, testing, quality assurance	Staying unseen, moving proceeds without a trace
<i>Tier two. The operator, not the crime. Scales and automates an assembly that already exists.</i>			
6 Accelerants	Anything that compresses the time or cost of the capabilities above	Optimization, efficiency	Not a crime ingredient. Removes the time limit and scales the assembly
7 Orchestration and autonomy	Agent frameworks, planners, tool routers, self-directed loops	Productivity	Not a crime ingredient. Removes the human so the assembly runs itself

The four outcomes downstream: fraud, deception induces a transfer; extortion, a coerced transfer under threat; theft, an unauthorized taking, usually unseen; and selling, the market where the other three cash out. Every convergence ships with its defense: a blindness overlay marking where existing coverage has no name for the harm, and a mitigation mapping naming what another domain already used as a remedy. Detection paired with a remedy, never a threat catalog. Sub-capability definitions are in the v2.2 reference of 12 September, which is downstream of this proposal and not part of it.

Appendix B. Limitations and Dual-Use Considerations

Limitations. The mapping is retrospective, and a scheme this rich can be fitted to a rich record after the fact; the answer is pre-registration, which is future work. The corpus is agent-authored and partly spoofed by its subjects. No base rate exists for what a full seven means, so the score has no denominator. The evasion row is the thinnest and is marked medium. The onboarding attempt produced no agreement statistic, because it ran for an afternoon against a same-night deadline with a definitional layer one day old and at the wrong unit of analysis; it is reported for what it does show and not as a measurement of the instrument.

Dual-use. A scheme that names which component turns a partial capability assembly into a self-running one is, read from the other side, a parts list. The mitigations are that every capability

named is already public and already exercised in documented incidents, that the coordination-surface class named in the forecast is already burned and published, that the instrument describes categories and roles and names no tools and no methods, and that its product is a warning paired with a starting remedy rather than a catalog of threats.

LLM Usage Statement

The lead author has a neurodevelopmental condition and uses speech-to-text (Wispr Flow) and an LLM as assistive technology to transcribe, clean up transcription artifacts, and format dictated analysis. The taxonomy, the mapping, the rating instrument, and the conclusions are the authors' own; both raters rated without an LLM. No factual claim in this report originates with an LLM.