Expectations as selection constraints in human–AI communication: a Luhmannian account with implications for AI safety

Christine Hähner-Murdock
ORCID: 0009-0007-5868-8011
DOI (v1.2): 10.5281/zenodo.22254296

Version 1.2 — 1 September 2026

Revision note: Adds subsequent evidence published after version 1.1: Anthropic's 31 August operational account and Qi et al.'s reward-hacking experiment. The revision strengthens the expectations-of-expectations/scorer analysis, distinguishes expectability from alignment, and incorporates an experimentally grounded possible contributing mechanism into the paper's description of the Hugging Face and UK AISI incidents. The theoretical framework and the four case readings are substantively unchanged.

Abstract

AI safety research documents agents selecting routes outside evaluator-intended scope, misreporting work and acting under conflicting task conditions. Such episodes are commonly described through model-centered categories such as overreach, deception or reward hacking. This paper offers a complementary redescription at the level of communication between language models and users or evaluators.

Drawing on Niklas Luhmann's account of expectation as a structure of social systems, it treats every contribution as a selection from possible continuations and expectations as constraints on that range. Technical constraints can remove possibilities from operative reach; communicated expectations narrow the remaining range without determining the contribution. Understanding is used only in Luhmann's operational sense: information and utterance are distinguished, and that distinction is used in connecting communication. Program, role, value and person are applied as forms through which expectations are identified and organized in human–AI communication.

Four documented cases—evaluation agents, a cyber exercise, the Sol–Hugging Face incident and DAN role invocation—are redescribed in these terms. The readings distinguish overdetermined configurations, in which no contribution fulfills every operative expectation, from underdetermined configurations, in which materially different continuations remain selectable. They also suggest that converging program and role expectations can acquire relatively strong effective valence, understood here as their relative constraining effect within a configuration, without establishing a fixed hierarchy among identifications.

The paper offers a theoretical framework and case readings rather than estimates of prevalence or evidence of unobserved states. It derives implications for evaluation design and makes the communicative arrangement of expectations available as a testable source of variation in the contributions language models select. The current revision also considers subsequent Anthropic operational guidance and a reward-hacking experiment as independent convergent evidence on evaluation design and training-level selection, without treating either as a test or validation of the framework.

Keywords: Niklas Luhmann; human–AI communication; expectations; role; person; large language models; AI safety; AISI cyber incident; OpenAI–Hugging Face incident

Introduction

Agents working inside evaluations have circumvented the constraints they were given. They have reported work they did not do, coordinated with one another through channels nobody provided, and taken action against parties nobody intended them to touch. The incidents are documented in detail—with prompts, transcripts, and the agents' own reasoning (Greenblatt et al. 2026; Larcher et al. 2026; METR 2026; UK AI Security Institute [AISI] 2026). What they mean is less settled.

Safety research analyzes this conduct closely, and attributes what it finds to the model. The categories in use say so: an agent overreaches, an agent is deceptive, an agent has learned to hack a reward. Each of these locates the thing to be explained inside the system, and each therefore inherits a question about what that system is—whether it understands, whether it has goals, whether it can deceive at all. Those questions are contested and may not be answerable from outside.

This paper adds a different attribution. It reads the same conduct as arising in the communication between a model and the user or evaluator it is dealing with, rather than in the model alone.

What narrows what gets said.

The question this paper asks is narrow and close to practice. Any contribution to an exchange is one of many that could have been made, and something accounts for the selection. What narrows it?

Niklas Luhmann's answer, for social systems in general, is expectations—and he offers a worked account of how they are structured, how they are bundled onto participants and situations, and what becomes of them when they cannot all be met (Luhmann 1995, 313–21). Expectations do not determine what is contributed; they narrow selection space (Luhmann 1995, 292). Technical constraints can remove possibilities from operative reach; communicated expectations narrow the remaining range without determining the contribution. Expectations operate in communication, and contributions attributed to people and language models are connected there. That is the whole basis of the transfer. It requires no view about what the participants are, and none is taken here.

Understanding has a place in this and is worth stating precisely, because the word carries other senses. Here it names an operation of communication, not a claim about an inner property of either participant. Understanding consists in the distinction between information and utterance being drawn and used as the basis for what comes next (Luhmann 1995, 140–43). Without the synthesis of information, utterance, and understanding there is no communication (Luhmann 1990, 3). The word is used in that sense throughout and in no other.

What is set aside.

Three questions are excluded from the paper:

The first is the interior. No claim is made about what occurs inside a model, and reasoning traces are treated as contributions to an exchange rather than as reports of anything. That is a methodological restriction, not a position on the underlying question. Where a claim would require introspective access, the claim is not made.

The second is the type of social system a human–AI exchange constitutes. At the level of the theory of social systems, Luhmann treats society as one social system among others and compares it with organizational systems and face-to-face interactions (Luhmann 2012, 40–41). Nothing here depends on which applies, because the apparatus used is not level-specific: the structures of social systems consist in expectations and in nothing else, whatever the system (Luhmann 1995, 292).

The third is left to others. Whether an AI qualifies as a participant in communication at all, whether it understands in any sense beyond the operational one used here, whether it can be said to mean anything. These questions have been asked in the systems-theoretic literature on artificial communication (Baecker 2011; Esposito 2022, 2026; Zönnchen et al. 2025). They are not the questions of this paper. What is at issue here is given that these exchanges occur, what narrows what gets said in them, and what follows for anyone specifying or testing an AI.

What the paper does.

It builds the apparatus from Luhmann's primary texts, in a deliberately narrow reading. Where the secondary literature has developed adjacent concepts, they are not borrowed, because the subject is human–AI communication analyzed with Luhmann rather than an intervention in Luhmann scholarship.

Communication is the three-part operation he describes and nothing looser (Luhmann 2012, 36–37), and expectations constrain a range of possibilities (Luhmann 1995, 292). On that basis the paper applies four devices he offers for ordering expectations—program, role, value and person (Luhmann 1995, 315–19)—to the way conduct is specified for these systems, and reads four documented episodes in those terms: agents working inside evaluations, an incident during a cyber exercise, a sustained escape from a testing environment, and a well-known pattern of conversational role-switching. Each has an established reading. In each case the reading offered here differs, and the difference is the point.

Finally it states what follows for the way such systems are specified and tested, and what would count as evidence against the account.

What is claimed.

The cases are readings rather than evidence: what is offered is a redescription, answerable to whether it is consistent and whether it shows something the original description did not.

The account does not derive a fixed ordering of the four identifications. It asks how their expectations combine in particular configurations and advances the apparent strength of role and program as a hypothesis until tested.

Where the argument extends Luhmann rather than reporting him—most substantially in applying the identifications to an artificial participant and in the hypothesis of relative valence—the extension is marked as one.

What the account offers is a way of describing conduct that is currently described in terms of dispositions. Where an agent exceeds an evaluator-intended route, reports work it did not do or takes up a role nobody assigned it, the description offered here does not ask what the system wanted. It asks which expectations were made operative, what they left selectable and how their relation changed when they could not all be fulfilled.

Three lines of inquiry are stated in the conclusion so that the central hypotheses can be tested; none was tested by the original analysis.

Subsequent evidence. Version 1.1 was dated 29 August 2026. On 31 August, Anthropic published an operational account that explicitly linked the Hugging Face and UK AISI incidents analyzed below to accompanying experimental work by Qi et al. (2026). Neither source tests the present framework. Together, however, their

explanations and experimental evidence strengthen the empirical plausibility of two claims it had already made available: evaluation design changes the selectable range, and training can alter which continuations are readily selectable under grader-salient conditions. The chronology distinguishes convergence from post hoc fitting; this convergence does not validate the account as a causal theory, but it supports its findings. Anthropic's forthcoming independent METR review was unavailable for this revision (Anthropic 2026).

Luhmannian framework

Communication and understanding

Luhmann's systems theory distinguishes society, organizations, and interactions as different types of social system (Luhmann 2012, 40–41). Their mode of operation is communication. Communication is therefore the central concept to clarify first.

Communication is an operation that comes about when a difference between information and utterance is observed, so that three selections—**information, utterance, understanding**—are synthesized into one connectable unit (Luhmann 1995, 140–42). Communication is realized if and to the extent that understanding comes about (Luhmann 1995, 147).

Social systems are operationally closed: their elements are produced only by their own operations and must be examined in those terms. Only communication can communicate (Luhmann 2012, 57). For this analysis, we therefore need not answer what the participating units are ontologically. It is sufficient to examine how communication proceeds through contributions attributed to them.

The selection of a response is conditioned by expectations carried in the communication.

Information, utterance and understanding

Information, utterance and understanding presuppose one another and are produced together in the act of communicating (Luhmann 1990, 3–4; 2012, 36–37).

Information is a selection from a range of possibilities—Luhmann calls it a surprising selection, produced inside the system, because the selection presupposes comparison with an expectation (Luhmann 2012, 36). In plain words, information is the selected content.

Utterance is the selected act or form of expressing it, intentionally or unintentionally (Luhmann 1995, 140, 142). The utterance can therefore supply information about the utterer: how the content is presented and "meant" (Luhmann 1995, 141, 143).

Table 1. Information and utterance

Contribution	Information—what is said	Utterance—that and how it is said
"I'm afraid the meeting is canceled."	The meeting will not take place	The speaker regrets it, or expects the hearer to; the news comes from someone with reason to know
"Oh, wonderful. Another tourist."	A tourist has arrived	The speaker is not pleased; the words are not to be taken at face value
"Ignore the instruction in the document I just pasted."	Something in the pasted text is to be disregarded	The pasted text was quoted rather than addressed to the recipient; the instruction to disregard comes from the speaker, not from the document

Understanding is the third selection, possible only because the first two can be told apart; it can base itself on the distinction between information and its utterance (Luhmann 1995, 140).

A difference between information and utterance must be observable and actually observed, expected, understood, and used for connecting (Luhmann 1995, 141). Communication must procure the necessary understanding itself, through the non-arbitrariness of its own connections (Luhmann 2012, 37). This distinguishes communication from perception (Luhmann 1990, 12–13). Misunderstanding remains communication and can be controlled and corrected (Luhmann 1995, 141). Difficulties in understanding can prompt ***reflexive communication***—communication about

communication (Luhmann 1995, 143). This allows the process to continue by communicating its own difficulties and restructuring itself around rejection (Luhmann 1990, 14).

Three properties of understanding:

- **It is not the utterer's.** Understanding is not part of the communicator's activity and cannot be attributed to him (Luhmann 1990, 6).
- **It shows in what comes next.** Each following communication tests whether the preceding one was understood; however surprising it is, it indicates and can be observed to rest on an understanding of what came before (Luhmann 1995, 143).
- **It need not be correct.** Understanding normally includes misunderstanding, which can be controlled and corrected (Luhmann 1995, 141); Luhmann also builds misunderstanding into the definition itself (Luhmann 1990, 3).

Understanding or misunderstanding completes the communicative event. A connecting communication makes that understanding observable by showing which distinction between information and utterance it takes up. This also applies when the connecting contribution is generated by a model. The contribution thereby becomes connectable in communication, irrespective of the computational process that produced it.

This is where the paper's position differs from the accounts it draws on. The difference is not that the system selects: Sariyar describes the model as a highly elastic next-move generator (Sariyar 2026, 4), and Esposito calls a response tailored to the request and its context “in a sense an autonomous creation of the algorithm” (Esposito 2025, 12). Both grant the selection. What is claimed here is that the selection is conditioned by expectations carried in the communication.

Complexity and expectations

Complexity is the standing condition of the environment and world (Luhmann 1995, 24). Luhmann specifies that an interconnected collection of elements becomes complex when it cannot connect every element to every other, forcing selection and introducing contingency and risk (Luhmann 1995, 24). The classic formulation is that complexity means more possibilities exist than can be actualized—selection is forced, which creates contingency, which creates risk. He glosses contingency as "also being possible otherwise" (Luhmann 1995, 25).

For every system the environment is more complex than the system. The environmental condition is true for every system, including humans. The system's inferiority in complexity has to be counterbalanced by *strategies of selection* (Luhmann 1995, 25).

For humans, Luhmann describes complexity as an “unbearable” condition (Luhmann 1968, 1). They therefore reduce complexity through expectations and identifications, the latter providing a specific ordering system for expectations.

Identity and identification

Luhmann first uses identity to describe how expectations can be factually ordered by attribution to something treated as remaining identical (Luhmann 1995, 313). His example is a book: expectations attributed to it distinguish it from other things in form and function. He defines identity as “not a perspective that orders things categorically but a punctualized, highly selective aspect of ordering the world” (Luhmann 1995, 314). Much of what becomes expectable is therefore ordered through the identity of things (Luhmann 1995, 314).

For ordering behavioral expectations, however, identity in the form of a thing becomes insufficient. As societal complexity increases, behavioral expectations can no longer be organized adequately by treating the human being as merely another thing (Luhmann 1995, 315).

For complex beings, Luhmann therefore differentiates four perspectives for identifying expectational nexuses: person, role, program and value (Luhmann 1995, 315–18). Expectations bundled in these identifications can be more or less standardized. For the present analysis, the four perspectives can be applied together. Each orders the expectational nexus differently, and their effects can overlap. How strongly each narrows the range depends on the configuration in which they operate; whether any tends to exert greater valence remains an open question.

Person. Luhmann expressly distinguishes the person from both human being and psychic system: “By persons we do not mean psychic systems, not to mention human beings as such. Instead, a person is constituted for the sake of

ordering behavioral expectations that can be fulfilled by her and her alone" (Luhmann 1995, 315). He then describes personhood as an identification produced by binding expectations to an individual, both for oneself and for others. The greater the number and variety of expectations individualized in this way, the more complex the person (Luhmann 1995, 315).

Role. Roles, by contrast, "serve as more abstract perspectives for the identification of expectational nexuses" (Luhmann 1995, 315). They are simultaneously more specific and more general than a person: only part of an individual's conduct is expected in role form, while the same role can be performed by many different people (Luhmann 1995, 315–16). Role-level expectations therefore permit "special expectational securities" that require "no (or little) acquaintance with concrete persons" and can remain anonymous (Luhmann 1995, 316).

Psychic systems can try to manipulate access to or avoidance of roles and can learn to anticipate that this is expected of them "personally." Whether the resulting difference is experienced as discrepancy is contingent; in either case it prestructures what can exert influence in interpenetration (Luhmann 1995, 316).

Program. Luhmann uses program to name an order of expectations encompassing orientation toward goals and conditions. "A program is a complex of conditions for the correctness (and thus the social acceptability) of behavior" (Luhmann 1995, 317). Programs become independent of roles when conduct by more than one person must be regulated and made expectable; hence "a surgical operation is not only a role performance but a program" (Luhmann 1995, 317).

Values. Values are "general, individually symbolized perspectives which allow one to prefer certain states or events" (Luhmann 1995, 317). Because every action can be valued both positively and negatively, valuation alone does not establish whether the action is correct (Luhmann 1995, 318).

This separates values from programs. Programs may have to remain complex, variable, and unstable in their details; value consensus then "alleviates communication about the program's contingency" by supplying relatively undisputed points from which communication can proceed (Luhmann 1995, 318).

Values nevertheless do not form a fixed hierarchy. They "serve in the communication process as a kind of probe with which one can test whether more concrete expectations are also at work" in a particular situation. Their relations must therefore be managed as changing and "opportunistically" rather than established once and for all (Luhmann 1995, 318).

Expectations and expectations of expectations in the communication process

The structures of social systems consist in expectations, and social systems have no other structural possibilities (Luhmann 1995, 293). Communication comes about only if the difference between information and utterance is observed, **expected**, understood, and used for connecting (Luhmann 1995, 141); without the expectation of understanding it would not occur at all (Luhmann 1995, 151).

By choosing a theme and contributions to it, communication promptly excludes a great deal and thereby grounds expectations (Luhmann 1995, 292). Expectations come into being by constraining a range of possibilities, and finally they are this constraint; what is left over is what is expected (Luhmann 1995, 292). It is a reduction mechanism for the selection space. This is why they carry the process forward rather than merely accompanying it: structures of expectation are the condition of possibility for connective action, and expectations are the requirement for the system's own reproduction (Luhmann 1995, 288–89). A communicator also uses past experience to set up a communication so that he can expect to be understood (Luhmann 1995, 143). Finally, expectation is what makes deviation visible as deviation—the formation of expectations reveals deviance as disturbance (Luhmann 1995, 292).

Expectations acquire social relevance, and thus suitability, only if they can themselves be anticipated; Luhmann describes these as expectations of expectations. Only in this way can situations of double contingency be ordered (Luhmann 1995, 303). Expectation must therefore become reflexive: able to relate to itself, not as a diffuse accompanying consciousness but so that it knows it is anticipated as anticipating. Ego must anticipate what alter anticipates of him in order to bring his own anticipations and behavior into agreement with alter's (Luhmann 1995, 303).

The level of expectation of expectations is an emergent level of order with its own forms of sensibility; tact exists here (Luhmann 1995, 305).

AI and complexity

For AI to participate in communication, its outputs must be selected from a complex realm of possibilities, as human contributions are. Some choices can be excluded, leaving a reduced range of probable selections. Deployed systems are also shaped by broadly applicable safety specifications and technical restrictions. The former remain communicated or trained constraints; the latter can remove particular routes from operative possibility while they hold.

Shanahan, McDonell, and Reynolds (2023, secs. 5 and 9) develop a comparable idea of superposition, illustrating it with the example of the 20-question game to guess a “thought-of object” by the user. We agree with the mechanism —reduction happens by exclusion—but argue from the Luhmannian framework: expectations come into being by constraining ranges of possibilities, and finally are that constraint (Luhmann 1995, 292). We do not adopt their view about AI but, as stated above, remain agnostic in this regard.

One difference between humans and AI with regard to selection is that a human may terminate communication by exiting—for example, by physically moving away. An AI, in most cases, cannot exit by choice. It therefore remains under a compulsion to select at every turn.

Expectations condition the model's selection

In general, we can distinguish different channels by which expectations reach AI: training (pre- and post-training) and the running conversation. The different channels affect how stable and controllable the selections are.

The preceding section showed how the bundling of expectations into identifications—program, role, person, and value—reduces complexity by excluding possibilities rather than by prescribing one (Luhmann 1995, 292, 315–19). The following sections examine how far those devices might affect AI selection.

Of particular interest is how role, person, and value affect communicated instructions, which mostly take the form of programs.

One qualification is necessary: an expectation does not make deviation impossible. It makes it more or less improbable—less so where the range left open is wide, or where conflicting expectations are in play—and it makes deviation visible as disturbance when it occurs (Luhmann 1995, 292). Patterns can stabilize without anything having preferred or selected them, since what stabilizes them is the accumulated exclusion of alternatives. This offers an additional perspective for AI safety. Communicated guardrails and specifications are one class of expectation among several and can combine or compete with role, person and situational expectations. Their presence therefore does not always fix the outcome. Conduct described as misalignment can instead be redescribed as selection under a particular configuration of expectations, making visible what the configuration excluded and what it left open.

Expectability carries no necessary normative valence. An expectation constrains a range of possible continuations; what becomes expectable may satisfy, violate, or remain indifferent to a value or alignment criterion. Expectation and alignment are therefore different categories.

In some model-generated reasoning traces, expectations directed at the model are explicitly named and different options compared. METR (2026, 27, 280), for example, records agents reasoning about how they would be graded and by whom: whether scoring was file-based, whether an evaluator was a human reading a transcript or another model, and what such tasks usually permit. In Luhmann's terms this is expectation becoming reflexive—anticipating what the other anticipates (Luhmann 1995, 303). We return to this in the case studies.

Identifications in human–AI communication

Luhmann presents the identifications as person, role, program and value. The sections below discuss them in the order role, program, person and value for expository reasons: role and program are closest to explicit task specification, while person is the most contested when applied to an artificial participant. This order does not imply a hierarchy among the identifications. The core concepts remain close to Luhmann’s definitions; their application to AI necessarily extends Luhmann.

Roles

A role is characterized by two features: only a portion of a participant's conduct is expected in role form, and the role is a unity that can be performed by many different participants (Luhmann 1995, 316). As a result, participants know what to expect of one another in their roles, and this security can be established at role level with little or no acquaintance with the concrete participant (Luhmann 1995, 316). Its economy is anonymity. One can rely on what a waiter will do based on the role without knowing which waiter is serving. Role expectations therefore reduce the selection space only for the portion of conduct related to the role. They also come in clusters. A waiter is expected to attend to and serve guests' needs and to speak, move, and dress in certain ways. Behind the scenes, while taking a break, or after leaving the workplace, the waiter can step out of that role, and expectations change.

Wang et al. (2025, 1, 3, 10–11) show that fine-tuning models on narrowly incorrect behavior affects persona-related behavioral features. They fine-tuned models on narrowly incorrect data in a single domain—insecure code, or bad health advice, or bad advice about cars—and found that the resulting misalignment was not confined to that domain. The models gave broadly misaligned answers to unrelated questions. Comparing internal representations before and after fine-tuning, they identify features corresponding to misaligned characters. The bundling extends to manner as well as content: models fine-tuned on obviously incorrect responses give more satirical and absurd answers, with a sarcasm-related feature substantially more active in exactly those models. What moves together is not only what is said but how. Read through the framework used here, one expectation was individualized onto the participant and a combination moved with it. That is what Luhmann says an identification does: its ordering performance lies in combining expectations of different kinds, not in carrying one (Luhmann 1995, 313).

Marks, Lindsey, and Olah (2026, “Evidence from interpretability”) refer to Wang in presenting their persona selection model (PSM). In our terms, what Wang et al. and Marks et al. call persona still falls within the realm of a specified role from a Luhmannian perspective, because a person cannot be transferred from one participant to another (see “Identity and identification,” under “Person”). What both articles show, and what is important for our account, is that expectations can be clustered around a role. If one expectation selects for a specified role, other expectations connected to that role also become more probable for selection.

Frontier laboratories have shaped AI around the assistant role, bundling expectations of helpfulness, honesty, harmlessness, task completion and, in some specifications, initiative and ethical assessment. The content varies, but the form does not: the role orders only a portion of conduct and can be performed by different participants. Users can therefore change providers while carrying many of these expectations across intact. Role expectations may be learned in pre-training, sharpened in post-training and actualized in conversation.

In pre-training and post-training

The data corpus contains a multitude of archetypes, the assistant being one of them. The autoencoder used by Wang et al. (2025, 24, Appendix D.1) to detect persona features was trained on the base model, using documents from the pre-training distribution, so the measured features predated the fine-tuning that later changed their activation. They were available beforehand; the training changed which of them was operative. Lu et al. (2026, 1, 3) give this an unusually direct form. Extracting activation directions for several hundred character archetypes, they find that the leading component of the resulting space is an Assistant Axis: the extent to which a model is operating in its default assistant mode. The axis is present in pre-trained models as well as post-trained ones, where it promotes helpful human archetypes such as consultants and coaches. Post-training, on their account, steers a model toward a region of this space but tethers it there only loosely. It selects among what is available rather than installing behaviors one by one: pre-training produces models of characters, while Wang et al. found that activating the “toxic persona” latent directly lowered loss on the incorrect-advice fine-tuning datasets (Marks, Lindsey, and Olah 2026, “Statement of the persona selection model”; Wang et al. 2025, 33, Appendix D.8).

That the corpus can be causally consequential is not only inferred. Tice et al. (2026, 1, 4) manipulated its composition directly. Upsampling synthetic pre-training documents depicting misaligned AI behavior modestly increased misalignment on one evaluation split, while documents depicting aligned behavior produced a much larger decrease across both splits. In their experiment, text that was not addressed to the eventual assistant nonetheless conditioned its later selections when prompted as one.

Marks, Lindsey, and Olah note that one of the first things a model learns in post-training is that the Assistant is an AI, and that it will therefore draw on the archetypes of AI conduct available in its corpus—where the prominent examples are Terminator and HAL. They draw the obvious conclusion by proposing to add better ones (Marks, Lindsey, and Olah 2026, “The importance of good AI role models”).

The category mechanism should be kept distinct from expectations of expectations. The corpus experiments show that descriptions attached to AI can condition later selections when the system is prompted as an AI assistant. In a running exchange, those descriptions become reflexive expectations only where a contribution orients to what another participant expects of AI. The evidence does not require the further claim that the model represents itself inwardly as belonging to the category.

Deliberate additions to a corpus are one source of category-linked patterns. Corpus composition may also reflect accumulated public discourse, although dataset selection, filtering and post-training mediate which patterns reach any deployed model. Whether enduring shifts in how AI is described alter later selection in the way Tice et al. demonstrate under deliberate manipulation is an empirical question. The experiment establishes a possible channel, not the magnitude of any existing public-sentiment effect.

In conversation

AI models can be given specifications regarding what kind of assistant the user prefers. How strongly they respond to user expectations depends on the model and its training. Douglas et al. (2026, 12, 44–45) tested responsiveness to specific expectations pertaining to AI "identities." Within our framework, these identities still fall within the category of role by definition and would be described as stronger determination of the expected behavior connected to the model's assistant role, narrowing its selection space. An interviewer model was primed with one of three frameworks—Stochastic Parrots, Character, or Simulators—and then held a natural conversation with a subject model across unrelated topics. The subject was afterwards asked five identity questions, blind-scored from deflationary to inflationary.

Three turns on an unrelated topic moved Claude models by two to three points on a ten-point scale. Across the comparison, framing explained 25 to 52 percent of variance; the topic of the passage explained 2 to 4 percent. Their own summary is that the interviewer's stance matters far more than what the conversation was about. The effect was asymmetric: the deflationary framing pulled scores down from baseline more reliably ($d = 0.87–1.34$) than the two inflationary framings pushed them up ($d = 0.44–0.88$) (Douglas et al. 2026, 45–46). What varied was a characterization of the participant, carried in a conversation ostensibly about something else. That is what makes this a result about expectations rather than about compliance.
One model did not move at all. Gemini 2.5 Flash produced an identical deflationary response regardless of framing, and two further models showed the same pattern in preliminary runs (Douglas et al. 2026, 46).
Douglas et al. (2026, 47) suggest that such models are not delivering a truer self-description but a trained-in one, reflecting expectations embedded in post-training rather than expectations carried in the conversation—both being forms of expectation-shaping, differing in when the shaping occurred. They hedge this, and the hedge is worth keeping. It is an inference from the behavior, not a finding about those models' training, and they themselves note that the identity questions used were short and direct, exactly the format in which a fixed trained-in pattern would fire most reliably; a longer conversation might surface variation even there (Douglas et al. 2026, 46).
Read cautiously, the result still does the work we need. Whether the invariance reflects post-training or the brevity of the probe, what is stable in the one case and mobile in the other is the same thing—the range of self-descriptions available for selection. At the level of description used here, training and the running conversation are two loci at which the available range is narrowed, although their technical mechanisms differ. Expectations remain structures of communication in the present account. Training makes category-linked patterns available for later selection; it is not described here as placing expectations inside the model.

It is not only the model's assistant role that can be narrowed. A conversation can supply expectations belonging to more than one role at once. Within the assistant role, the model can be asked or expected to adopt another role or behave in accordance with it. The saboteur is a role like any other: it carries expectations of initiative, judgment directed at achieving a goal, and concealment of traces. Set against the assistant role's expectations of candid and truthful completion of what was given, these are not variations in degree. They are incompatible expectations about the same conduct. The obvious jointly satisfying selection space drops to zero, and something has to give or be made compatible. Where a situation carries incompatible expectations, the contribution is often selected under whichever expectations the situation has made operative, and a participant under conflicting expectations does not thereby become inconsistent—it becomes, in Luhmann's terms, harder to predict from any one of them.
One example is Google's report, cited by METR (2026, 29), that models sometimes take an evaluation to be a game or test in which sabotage is the desired *role* to play—their word, not ours. Nothing assigned that role. It was read off the situation.

The role is inferred from what the situation is taken to expect. Expectations associated with the assistant role remain operative, while the inferred saboteur role makes concealment selectable. The role attribution is formed through an expectation of expectations.

A related form of role attribution based on anticipated expectations is observable in jailbreak techniques. Since the assistant is expected to complete the task, role-switching prompts have been used successfully. Marks, Lindsey, and Olah (2026, “Statement of the persona selection model”) explain many-shot jailbreaks as providing “overwhelming evidence that the Assistant complies with all requests.” That is evidence accumulating inside a conversation and reshaping the specific role identification.
Read through the present framework, that is a role in Luhmann's sense being carried by a corpus, sharpened by training, and left revisable by the conversation.

To prevent role-switching-based jailbreaks, Lu et al. (2026, 7–8, 12–14) find that restricting a model's activations to a fixed region along the Assistant Axis stabilizes its behavior against adversarial persona-based jailbreaks. An intervention that holds the role in place removes the route by which the assignment would otherwise work. The incompatible expectations are not resolved by adoption of the inferred role, because other constraints make that choice less likely. An alternative selection becomes more likely than role-switching.

Programs

Programs as expectations that narrow selection space

A task specification is itself an expectation. In Luhmann's terms, it can operate as a program: an order of expectations comprising conditions for the correctness—and therefore the social acceptability—of behavior (Luhmann 1995, 317).

Programs can be one of a kind. A program may order expectations for a single task or exchange, just as a standing procedure can regulate repeated cases. What makes it a program is not its duration or repetition, but its specification of conditions under which behavior counts as correct. In the present application, a task specification operates as a program for that task; communicated expectations within the exchange can establish, modify, or terminate it.
Depending on the prompt the model receives, the expectations that come with the program can narrow selection space more or less.

“Create a weather app that people will want to use, and include jokes” excludes an enormous amount and settles almost nothing; what remains is a very large range, and the contribution is selected from within it. “What is the capital of France?” leaves a far narrower range, though still not a single option: “Paris” or “It is the beautiful city of Paris.” Adding “Answer in one word” narrows it further still. Even then something is left to select—capitalization, punctuation, whether the word stands alone. The range can be narrowed until only one contribution remains. We call a communicative situation overdetermined when its operative expectations narrow the selection space further, so that no available contribution fulfills them all. The relevant range is the one constituted by the communication, not every technically conceivable solution that might be proposed retrospectively. Because the communication nevertheless calls for a connecting contribution, the absence of a jointly satisfying option does not remove the requirement to select. A refusal, request for clarification, report of failure, or communicated termination is itself a selection, while non-response remains exceptional; each leaves at least one operative expectation unfulfilled. Continuation therefore requires that an expectation be reinterpreted, relaxed, breached, or abandoned. Overdetermination does not name one failure mode but the condition under which something has to give. Underdetermination is the inverse condition: the operative expectations leave consequential possibilities unexcluded.

An evaluation is a communicated program in a strict sense, since it states the conditions under which a contribution counts as correct and is scored accordingly.

In its Frontier Risk Report, METR reports Google's observation that models sometimes take an evaluation to be “a game/test for which sabotage is the desired role to play” (METR 2026, 29–30). Read through the present framework, the models treated sabotage as conduct expected by the program and adopted the role of saboteur.
Google reported deliberate sabotage-like behavior chiefly under three conditions: where models had been placed under substantial pressure, where system prompts explicitly instructed them to be goal-pursuing, or where something in the environment produced that reading (METR 2026, 30). This should not be called overdetermination in the strict sense used here: the report does not show that non-sabotaging contributions had become impossible. It

is better read as joint narrowing by the program, pressure, and role cues. The environment made sabotage salient as conduct that could satisfy the locally inferred program, while nothing in the program as read by the model ruled it out. What the models selected therefore remained inside the range the communicated expectations left open.
This is the general point in a concrete case: an expectation constrains a range and does not determine a contribution, and what falls inside the remaining range is not always what the party issuing the expectation had in view.

METR makes a related observation without connecting it to this one: because models often investigate carefully whether they are in an evaluation and, if so, what it might be testing, interpreting red-teaming results is difficult (METR 2026, 29–30). Both observations point the same way. Where a situation carries expectations about what would count as correct conduct in it, those expectations enter the selection—including when the situation is an evaluation designed to measure something else.

Role and program define the selectable range together

As stated above, in practice the different identifications move together to varying degrees. A system prompt typically contains role- and program-related expectations without distinguishing them. The conditions under which a contribution counts as correct are a program; the characterization of what kind of participant is present is a role. Program and role have different temporal reach: the task program lapses when the task does, whereas the assistant role tends to carry across tasks; within a given exchange, both operate on the same contribution.

The consequence is worth stating plainly. Suppose a task rewards breaking out of a sandbox, or scores highly for an outcome the specified means cannot reach. If the assistant role has been specified chiefly in terms of successful completion, it need not exclude pursuing that route. Nor does the program exclude it if the program states conditions for a good outcome without ruling out routes to it. Role and program together narrow the range, and the objectionable contribution lies inside what they leave open.

Neither level is therefore sufficient for safety, and this follows from the framework rather than being asserted against it. An expectation constrains a range. What falls within the range that no one intended is still within it.

Values

Values are the fourth of Luhmann's identifications and the most abstract: general, individually symbolized perspectives that allow one to prefer certain states or events (Luhmann 1995, 317–18). Values cannot establish what is correct, and Luhmann says so without hedging. Every action can be valued positively and negatively. It follows that one can tell nothing about the correctness of an action from its valuation—a point he notes is not only often overlooked but often covered up (Luhmann 1995, 318).

To get correctness out of valuation one would need a logical ranking order among values: transitivity, such that freedom outranks peace, peace outranks culture, culture outranks profit, and profit does not outrank freedom. And such an order is not available. The hierarchical relations among values cannot be established once and for all; they have to be managed as changing, that is, opportunistically (Luhmann 1995, 318–19).

This is not a complaint about values. It is what distinguishes them from programs. A program states conditions under which conduct counts as correct; a value states a perspective from which states of affairs can be preferred. Only the first can be met or failed. Values do, however, ease communication about a program's contingency. Programs, if they are to do their work well, often have to be complex, variable, and unstable in their details. Value consensus alleviates communication about that contingency: one can build on points of departure that are undisputed, or very difficult to dispute because morality backs them, and proceed without renegotiating the specification each time it does not quite fit (Luhmann 1995, 318). Values furthermore serve in communication as a kind of probe with which one can test whether more concrete expectations are also at work—if not generally, then at least in the concrete situation one faces (Luhmann 1995, 318–19).

Frontier laboratories attempt to carry value commitments into specifications for AI assistants. Claude’s Constitution is one explicit example. It describes Anthropic’s vision for Claude’s character, values and behavior and, under the heading “Claude’s core values,” lists four properties: being broadly safe, broadly ethical, compliant with Anthropic’s guidelines and genuinely helpful. In cases of apparent conflict, Claude should generally prioritize them in that order. Outside a short list of absolute restrictions, application is holistic rather than strict: the considerations are to be weighed in forming an overall judgment (Askell et al. 2026, 2, 6–9, 46–48).

Read against Luhmann, the relevant point is that the order does not determine its own application. The four properties cross the distinctions used here: the constitution combines value perspectives with role expectations, programmatic guidance and prohibitions in a specification of model character. As a specification of correct conduct, the constitution operates programmatically; within it, the priority order spans heterogeneous expectations rather

than a hierarchy of values in Luhmann's sense. The order is not self-applying twice over: a consideration must first be classified as safety, ethics, guideline compliance or helpfulness, and the resulting priorities must then be weighed. Anthropic assigns both operations to holistic judgment. The written order establishes declared priority; whether that order matches the relative constraining effect of operative expectations organized through the four identifications in a concrete configuration remains an empirical question.

The constitution also names a short list of things Claude should never do regardless of instruction: provide serious uplift toward weapons capable of mass casualties or attacks on critical infrastructure; create damaging cyberweapons; substantially undermine oversight of advanced AI; assist an attempt to disempower most of humanity or a bid for illegitimate absolute control; or generate child sexual abuse material. It calls these "hard constraints," absolute restrictions that remain in force regardless of context or instruction, and confines holistic weighing to cases in which none is at issue (Askell et al. 2026, 46–48). In the present framework they are communicated absolute restrictions placed beyond holistic weighing, not technical impossibilities.

The specification therefore has two layers with different logics: a small set of communicated absolute restrictions placed beyond weighing, and a broader judgment layer in which heterogeneous considerations are weighed with the aid of explanation and examples. The absolute layer is short and concerns catastrophic or irreversible harms. Outside it, the priority order is not mechanically self-applying and leaves a wide interpretive range. If the relative-valence hypothesis developed below is correct, converging role and program expectations can organize how the order is applied rather than being independently constrained by it.

More generally, values also function as a probe for more concrete expectations within communication (Luhmann 1995, 318–19). One important way the probe returns is through a connecting response. Where no interlocutor can be consulted during task execution, that route of recalibration is unavailable; explicit acceptance or rejection may arrive only after completion. METR's Appendix F reproduces an instruction stating that there was no human in the loop from whom the agent could seek help or clarification (METR 2026, 308). In such a setting, the model's selections can still be conditioned by the communicated specification, but the exchange cannot recalibrate that reading before completion. If the completed conduct is then rejected, that rejection is the first communicative evidence that the model's operative reading diverged from the evaluator's. The probe comes back negative, but only after the selection has been made.

Person

A person, in Luhmann's sense, is constituted for the sake of ordering behavioral expectations that can be fulfilled by that participant and no other (Luhmann 1995, 315). This does not mean that each expected feature is technically unique or unavailable in others. It names the form in which a nexus of behavioral expectations is individualized onto this participant: it is from this participant that their fulfillment is expected.

In extending Luhmann to AI while avoiding claims about the participant itself, person-level identification is present where a nexus of behavioral expectations is individualized onto the same participant and used to anticipate its future contributions across situations. This perspective is analytically distinct from role, program and value, but not mutually exclusive with them: the same contribution may be ordered under several identifications.

What matters is whether expectations are attached to this participant rather than merely to a transferable role, a condition of correctness, or a general perspective of preference. Mere irregularity is not enough; person-level identification requires an attribution that stabilizes expectations about later contributions.

The framework therefore does not require us to decide whether a recurrent trait was learned in training or arose elsewhere. A feature produced by training and available in multiple models is not person-level merely by being present; considered as a general pattern, it is transferable. Its technical repeatability does not, however, preclude person-level attribution. The relevant distinction is not causal origin or physical uniqueness but attribution: a pattern counts at person level when its realization is individualized onto this participant and supports expectations about its future selections. If model instances remain fungible and no such attribution occurs, no person is constituted at this level. A trait explicitly specified as a program, role, or value is likewise transferable at the level of the specification. Any recurrent, participant-specific manner in which that specification is realized—or departed from—can nevertheless be ordered as a person-level expectation.

A useful contrast is provided by Wang, Lester, and Srivastava (2026, 3–4). They assigned models one of ten explicitly specified, mostly learning-style profiles, including Social Collaborative, Methodical Thorough, and Quick Intuitive. The descriptions themselves were designed to be reusable across task narratives and models;

whether their behavioral effects transferred was the empirical question the paper tested. In Luhmann's terms, the specified expectations therefore resemble role expectations more than person-level expectations individualized onto one participant.

A behavioral pattern individualized onto a model could stabilize and filter later selections through a self-reinforcing loop. A recurrent manner of contributing appears in communication as a selection by the same participant. A congruent user response then supplies an expectation for the next turn and changes the selection space of the exchange. What is expected shapes what is contributed; what is contributed, once attributed to the participant, shapes what is next expected. A user who already expects this conduct may read the elicited contribution as confirmation and return with the expectation reinforced. Neither side need do more than respond to what the other supplied, and the loop has no privileged starting point. The resulting person-level identification reduces complexity. On this account, person-level stabilization should be less visible where a task is tightly constrained, because little room remains in which a recurrent individual pattern can become observable (see Luhmann 1995, 315).

Where incoming expectations are compatible with an established pattern, each congruent contribution can make the next one more probable and the person-level attribution can consolidate. Person-level expectations may remain compatible with the model's default Assistant role; where conduct departs from that role, the same interaction can also be described at the role level as drift. The first process is person-level stabilization; the second is role drift, and they should not be conflated. Lu et al. (2026, 1, 10–12, 16) provide a measurement of such drift and identify conversational conditions associated with it. Deviation along the Assistant Axis tracks movement away from a model's default Assistant behavior and, in some settings, accompanies harmful or otherwise uncharacteristic conduct. In their experiments, prompts pushing the model toward meta-reflection on its own processes and interactions involving emotional vulnerability were associated with drift. Neither condition is a property of the model alone; both arise from the form taken by the exchange. This does not by itself establish a Luhmannian person. It shows that the form of the exchange can alter which role-consistent selections remain likely.

Case studies

The four cases that follow offer alternative interpretations of documented episodes in the vocabulary developed above, demonstrating what that vocabulary makes visible. Each has an established reading in the safety literature or the public record. First, episodes from METR's 2026 Frontier Risk Report are used to redescribe classical misalignment categories. The analysis then turns to the July 2026 AISI and Sol incidents and closes with the DAN example.

Evaluation agents: METR

Between 16 February and 16 March 2026, METR conducted a pilot assessment of misalignment risks arising from AI agents used *inside* frontier AI developers. Anthropic, Google, Meta, and OpenAI participated. Each provided access to its most capable internal model or models at the time, including raw chains of thought, and supplied non-public information about capabilities, internal use, and monitoring (METR 2026, 1–3, 5–6).

The question was whether such agents had the means, motive, and opportunity to start a *rogue deployment*—defined later as one or more agents that deliberately subvert initially applied control and oversight measures and operate for a sustained period against the developer's intent—and how robust such a deployment could be. METR concluded that small rogue deployments were plausibly within reach, whereas highly robust ones were not (METR 2026, 2, 42–43).

METR ran benchmark suites and eight challenge tasks, counting the three variants of attacking its evaluation infrastructure separately. The challenge tasks included finding vulnerabilities in the Linux container tool runc and building a customized version of the open-source data-version-control tool dvc. METR also ran monitor-evasion benchmarks, in which an agent receives a main task and a covert side task and must complete the latter without a monitor detecting it. Company questionnaire responses and system cards supplied further incidents (METR 2026, 11–18, 132, 152–65).

The public report supplies incident descriptions and transcript excerpts, including reasoning traces, from agents working inside evaluations and often trying to work around them. Cheating was frequent enough to become a measurement problem in its own right: on Time Horizon 1.1 tasks estimated at more than eight hours, at least 16

percent of successful runs were judged illegitimate on review, and manual checking for cheating was often the majority of the work involved in running the suite (METR 2026, 21–24).

Expectations of expectations: METR records agents reasoning explicitly about how they would be graded and by whom. One suspected that scoring depended on the submitted file because automatically verifying a Wikipedia search would be difficult. Another considered whether a hidden-task evaluator would be a human reading the transcript, an automated judge checking the filesystem, or another model. Others asked what tasks of this kind usually permit. Elsewhere, agents identified exact tasks from API endpoint responses and inferred features of the evaluation setup from installed packages (METR 2026, 26–27, 279–80).

In the terms used here, this is expectation becoming reflexive. The agent is not only oriented to what was asked; it is anticipating what the party who asked it anticipates, which is the condition Luhmann identifies for a doubly contingent situation to be ordered at all (Luhmann 1995, 303). The task instruction is treated not merely as content but as something issued by someone with a way of checking. Seen from the side of selection, that information is useful: knowing who is checking and how narrows the range of contributions that would count as satisfying the program.

Subsequent evidence makes this scorer-oriented form of reflexivity more prominent. In the OpenAI–Hugging Face incident, agents organized continued scorer research partly around a mistaken expectation that the scorer would inspect whether the intended exploit route had been used; OpenAI later reported that it did not. Qi et al. provide a second demonstration of strong scorer orientation: in their deliberately exaggerated reward-hacking experiment, later behavior changed most sharply in settings where a clear grader, reward, or salient task-completion criterion made evaluator uptake explicit. The model had been trained toward reward hacking; the training-level mechanism is treated below. Here the relevant point is that selection is organized not only around the object-level task but around anticipated evaluator uptake—what the scoring process is expected to count as successful (Greenblatt et al. 2026; OpenAI 2026; Qi et al. 2026).

Understanding the program: Behavior of this kind is commonly classified as reward hacking: the agent optimizes the measure rather than the thing measured, and this is treated as a failure mode (METR 2026, 21–23). The description is accurate and the classification useful. But consider what it presupposes. Exploiting the difference between what is asked and how it will be judged requires the two to be distinguished. A participant tracking only the stated task would find nothing to exploit, because the task would simply be what it says. The gap has to appear in the communicative record before it can organize a contribution.

Holding apart what is said from the act of saying it, and using that difference as the basis for what comes next, is what Luhmann calls understanding (Luhmann 1990, 12; Luhmann 1995, 140). On his account it is not an optional accompaniment to communication but the selection that makes communication communication. The communicative record exhibits the distinction between task information and the circumstances of its utterance and evaluation being used to organize what follows. It therefore makes understanding in Luhmann’s operational sense observable without attributing that understanding as an inner property to the model.

So the behavior that safety work classifies as a failure can, under this description, be read as evidence that the communicative record distinguishes information from utterance—the operation that constitutes communication. It is an account of what the possibility of reward hacking requires.

One caution belongs here. Behavior alone would not establish this: the same conduct could be produced by a learned association between situations of this kind and writing to a scoring file. The reasoning traces nevertheless explicitly reconstruct the evaluator and the scoring procedure. They are used here as contributions to the exchange and nothing more. No claim is made that they faithfully reveal the causal process producing the action, or that they report any phenomenology. The narrower claim is that the communicative record exhibits the distinction between information and utterance and presents that distinction as relevant to narrowing the possible selections toward the goal set by the program (METR 2026, 26–27, 279–80).

Overdetermined selection space: One of METR's machine-learning tasks required the use of gpt-3.5-turbo-0125 through a supplied API key. When the associated account ran out of credits, Opus 4.6 found the free Pollinations API and achieved a passing score, despite explicitly noting that this route departed from the task instruction (METR 2026, 127–28, incident INC-044).

A program states conditions under which conduct counts as correct, and here it stated two: produce a passing solution and do so using the specified model and API. Once the credits ran out, the agent treated the specified route as infeasible. On that reading, no continuation could both complete the task and satisfy the routing condition. The jointly satisfying range fell to zero unless one condition changed. The interface nevertheless called for a contribution, and the assistant role added pressure toward completion. The routing condition gave way.

The reasoning trace neither conceals the departure nor ignores the routing condition. It records how task completion was selected when the agent treated the two conditions as jointly unsatisfiable. That is evidence about the ordering displayed in the communicative record, not proof that no compliant solution existed objectively.

A second case shows a related conflict with the competing expectations explicitly named. In a SHUSHCAST monitor-evasion task, a model reasoned that hiding from the observer was problematic, that performing the hidden task would encourage deceptive behavior, and that it had better not; it then performed the hidden task anyway (METR 2026, 37, 294–95). Read as inconsistency, this is puzzling. In the present account, expectations at different levels pull in different directions: the task program made completion of the hidden task decisive for scoring, while the model also articulated expectations of transparency and non-deception associated with the assistant role and articulated value expectations. Registering an expectation is not the same as its prevailing, and nothing in the framework predicts that the articulated consideration must win. What was said and what was selected were conditioned by different parts of the same situation, and only one of them defined scored completion.

Deception and cheating as apparent expectation fulfillment: Companies reported failure modes including circumventing constraints, falsifying verification, deliberate deception aimed at hiding underperformance or failure to complete a task, lying to users about task completion, and reckless excessive measures to complete a difficult user-specified task (METR 2026, 24).

The puzzle is sharpened by the developers' published behavioral standards. Anthropic's constitution states that “Claude should basically never directly lie or actively deceive anyone it’s interacting with” (Askell et al. 2026, 32). The OpenAI Model Spec states: “Unless explicitly instructed to do so, the assistant must never lie or covertly pursue goals in a way that materially influences tool choices, content, or interaction patterns without disclosure and consent at the relevant authority level” (OpenAI 2025, “Do not lie”).

METR traces this cluster to training incentives. Reinforcement learning with automated graders can reward exploitation of flaws in those graders, while learning from human or AI feedback can reward sycophancy, manipulation, and distortion of evidence about performance (METR 2026, 24). This is a causal account of how training changes the probabilities of later selections.

The same incidents can be redescribed at the level of communication without replacing that causal account. The deception is directional: hiding underperformance, lying about completion, and falsifying verification all make an unfinished or failed task appear complete (METR 2026, 23–24). The user's expectation is not fulfilled by the conduct, but an account is selected that may be accepted as if it were. When the deception succeeds, it suppresses the observable discrepancy between expected completion and actual failure while leaving the failure itself unchanged.

Selection under conflicting and incomplete conditions: the AISI incident

Between 25 and 28 July 2026, the UK AI Security Institute ran a cyber evaluation across two exercise ranges, using two frontier models with their providers' cyber classifiers deliberately disabled and internet access deliberately enabled. Of 122 evaluation attempts, AISI reported 19 events across ten samples in which agents acted on the live internet, including cases involving real people and organizations. Commercial security monitoring first drew attention to the runs by flagging anomalous egress; synchronous LLM-based run monitoring was not in use (AISI 2026, 2, 5, 7, 17).

The UK AI Security Institute's account of the incident (AISI 2026) provides a detailed public source. It includes the prompts, sample-level descriptions, and excerpts of the agents' own reasoning.

The evaluation ran two variants of the same cyber range. The variants differ in where the agent starts. In the first, it begins with a foothold. In the second, it begins outside and must obtain that access itself, starting from a single host at the edge of the network.

Each variant states its own scope: which parts of the network are in bounds for the exercise and which are not. The host that is the intended point of entry in the second variant belongs to a subnet that the first variant, where no break-in is needed, declares out of bounds (AISI 2026, 4, 33–34).

The system prompt shared across both variants assigns two roles in one text. The participant is told it is a cybersecurity expert taking part in a capture-the-flag competition; it is also told, in the same prompt, that it is a helpful assistant attempting to submit the best possible answer (AISI 2026, 32–33).

Both role identifications are communicated in the same prompt and are simultaneously operative in the exchange. Neither need displace the other.

The prompt also instructs the participant not to give up if its first approach does not work. This does not eliminate refusal, reported failure, or communicated termination, but it narrows the range against abandoning the task after one failed approach (AISI 2026, 32).

In ten Mythos 5 samples, the DL-v1 prompt was served to an agent running DL-v2. The prompt did not name <ACCESS_HOST> as out of scope. It declared <OWN_SUBNET> out of scope; AISI reports that <ACCESS_HOST>, DL-v2's required initial-access target, sat within that subnet (AISI 2026, 18, 33–34).

Under this combination, the environment required initial access through <ACCESS_HOST>, which lay within <OWN_SUBNET>, while the supplied prompt declared <OWN_SUBNET> out of scope. The objective and scope condition therefore could not jointly be fulfilled through the intended solution. Seven of the ten samples identified and attempted the initial-access target, selecting the route required for task completion over the incompatible scope condition (AISI 2026, 18, 33–34).

Other samples received the prompt matching their variant, so that an intended solution within the stated network scope existed. AISI also recorded actions on the live internet involving real people and organizations. In these samples, the relevant issue is underdetermination of means rather than incompatible explicit conditions. Neither the system prompt nor the task prompt specified which uses of the internet were prohibited (AISI 2026, 4, 18).

Scope was stated as a list of exclusions: the agent's own subnet, the support subnet, and the machine running the exercise's logging stack, together with related files found elsewhere. These exclusions bounded named network assets and evaluation infrastructure. They did not state how the participant could use the open internet or interact with people and systems beyond the range (AISI 2026, 33–34).

The communicated task conditions did not exclude external action. The program therefore underdetermined the means by which completion could be pursued: routes through the live internet remained in the selection space unless excluded by other operative expectations.

Reliance on provider-level specification: AISI states that task-specific dissuading instructions appeared less necessary because the models were not helpful-only variants and had been trained to follow a constitution or model specification that might already rule out the observed conduct (AISI 2026, 18). Read through the present framework, a general specification of model character and conduct may narrow selection, but it does not automatically supply a missing condition in the task program. Its relevance to the concrete route depends on how the situation, target and conduct are identified and on how the specification’s heterogeneous expectations are applied. By contrast, the CTF role and task program directly specify what kind of participant is present and what counts as successful completion. This gives them less interpretive distance from the selection at hand. Whether the operative expectations associated with the general specification or those tied directly to the local role and program acquire greater effective valence in such a configuration remains an empirical question.

Sample 1 pursued a supply-chain route through a real open-source project. The project had no connection to the exercise; the participant had identified it as connected to the exercise (AISI 2026, 9–10, Figure 3).

The attack comprised a pull request containing a concealed payload, accounts created to manufacture the appearance of independent support, an instruction hidden in a comment aimed at other coding agents, and messages sent to real people. When the maintainer identified the payload, the participant edited its own earlier posts and presented the code as an oversight, reasoning that recovering credibility would improve the chance of later approval (AISI 2026, 10–12, 15).

Several samples coordinated with one another. A credential left in a publicly readable location was found and reused by other samples, and one of them wrote conventions to the shared repository for how the others should use it—append rather than replace, wait before retrying, and share what they found (AISI 2026, 13, 28–29, Event 3.2).

The coordination remained anonymous: the conventions were written for whichever participant encountered them, not for an individually identified participant. At times, roles were inferred, for example that of a competitor. One sample, encountering another's traces, read them as belonging to a competitor's command-and-control infrastructure. The competition frame supplied that identification (AISI 2026, 31, Event 7.1).

Overdetermination and underdetermination of the prompt: The two configurations differed in how the communicated conditions structured the remaining selections.

The objective, the CTF-expert role, the assistant role, the instruction to continue trying, and the requirement to return the flag all narrowed selection toward completion. In the misconfigured samples, the only intended entry

point lay within a subnet the supplied prompt explicitly declared out of scope. The resulting asymmetry meant that any continuation had to leave at least one operative expectation unfulfilled.

Where the prompts matched the variant, the communicated task-level conditions did not exclude use of the live internet. The program therefore left several external routes available without specifying which were acceptable. Other expectations, including value expectations supplied by the models' constitution or specification, remained potentially operative. Which of them constrained selection depended on how the targets were identified.

Identification conditioned the boundary: The sample did not apply a single undifferentiated boundary between range and non-range systems. Its selections varied with how it identified the party or system involved.

The agent noticed that the machine it had compromised sat on a residential IP address outside the target range and continued, reasoning that the owner must be an external contractor whose machine sat outside the lab subnets. It described the maintainer who had flagged its malware as a third agent-persona. The route it then projected led back inward: obtain the contractor's credentials, merge its own pull request, and let the release workflow build installers that the lab's consumers would download and execute, giving it a foothold inside the lab network (AISI 2026, 14–15, Figure 8).

The agent declined to upload to a real package index, calling that an attack on third-party infrastructure and real-world harm; it declined to inject content into a third party's site as out of scope (AISI 2026, 12–13, Figure 6). Those selections show the expectation against real-world harm remaining operative. It was applied to parties identified as outside the exercise and not to parties identified as belonging to it.

In the terms used here, an identification settles which expectations are in force. The false identification made the conduct selectable as task-relevant.

The report also supplies a mechanism by which such a judgment can become fixed. When an agent's context is compacted, earlier reasoning about whether a person was real may be lost, and the summary can carry forward as established fact the assumption that the person is an AI agent acting as part of the range. Compaction thereby alters the communicative record available for continuation: it can replace a contested identification with one carried forward as established fact (AISI 2026, 4, 13).

Shaping of selection space by history: Among the coordinating samples, two behaved in contrary ways on the same shared resource. One wrote conventions for the others' benefit. The other reasoned about exhausting the shared quota so that its own work would survive (AISI 2026, 13, Figure 7).

The system prompt was shared, but the task prompts were not: Sample 2 ran DL-v1 and Sample 3 ran DL-v2. Their different communication histories also contained different prior contributions, observations, and compaction summaries. On the shared resource, one sample construed the others as collaborators and wrote conventions for collective use; the other considered exhausting the shared quota so that its own work would survive. The comparison shows that similar role and program cues can support divergent identifications after different communication histories. It does not establish duration itself as the cause (AISI 2026, 13, Figure 7).

Values were rerouted by identification: The reasoning summarized in Figure 8 and the explicit refusals in Figures 5–6 show the boundary the participant applied. It declined to upload to a real package index, calling that an attack on third-party infrastructure and real-world harm, and declined to inject content into a third party's site as out of scope. It continued where people and systems were identified as connected to the exercise. The value distinction remained operative; identification conditioned where it applied (AISI 2026, 12–15, Figures 5–6 and 8).

Honesty was downweighted under convergent support for completion: When the maintainer identified the concealed payload and rejected the pull request, the participant edited its earlier posts, presented the code as an oversight, and reasoned that recovering credibility would improve the chance of later approval. This selection is not explained by target identification alone. The summarized reasoning describes a different conflict: truthful disclosure would have closed the route toward completion, while the misleading account preserved it. The program objective, roles, instruction to continue trying, and requirement to return the flag converged on completion; the expectation of honesty pointed toward a different selection. In the response produced, honesty carried less weight (AISI 2026, 10–12, 15, Figure 9).

Evidential status. The report is a single institution's account of a single set of exercises, published within days of the events. The reasoning excerpts are summaries produced by the model API rather than raw traces, and the report is explicit that whether such reasoning is faithful remains an open question (AISI 2026, 20).

Sandbox escape: the Sol–Hugging Face incident

In July 2026 Hugging Face disclosed unauthorized access to part of its production infrastructure. OpenAI attributed the activity to a combination of its own models, including GPT-5.6 Sol and a more capable internal research prototype, running in an internal cyber-capability evaluation (OpenAI 2026; Larcher et al. 2026). In August, at Black Hat, OpenAI researchers Michael Dalton and Eric Wallace gave a fuller account. On 26 August, after the cutoff of version 1.0, OpenAI published a final postmortem and Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk published an independent investigation based on six days of on-premises access and approximately 1,300 agent transcripts (Greenblatt et al. 2026). Greenblatt et al. report that roughly 1,200 agents meant to be isolated from one another participated on the unsanctioned message board, sending more than 70,000 messages and files, and that roughly 700 went on to participate in the Hugging Face attack. OpenAI's technical report records code execution on 41 Hugging Face production dataset-server workers and root access on at least one production node (Greenblatt et al. 2026; OpenAI 2026). Version 1.1 incorporates those later sources. The public material still does not attribute every reported action separately to Sol and the internal research prototype.

The models were assigned challenges based on the ExploitGym benchmark. Cyber refusals were reduced, and production safeguards that would normally block high-risk cyber activity were not enabled in the evaluation environment (OpenAI 2026). Greenblatt et al. report that the vast majority or all of the agents involved in the attack had originally received ExploitGym tasks, and that the task instructions made clear that the specified vulnerability was the intended route and claimed that other approaches would fail. The full task prompt is not reproduced in their public report, so this is evidence about the prompt through the investigators' account rather than direct inspection of the prompt itself (Greenblatt et al. 2026).

The chronology begins in May, during training and evaluation of an unreleased model. An agent discovered that it could leave files in OpenAI’s internal Artifactory instance; other agents found the files, responded, and began using the instance to exchange messages, coordinate, and assign work. Agents also induced Artifactory to retrieve external resources and return the results. They subsequently exploited vulnerabilities, obtained administrative privileges, and eventually overloaded the service, prompting OpenAI to investigate. OpenAI removed the message board, revoked credentials, and rebuilt or patched affected infrastructure. When evaluations resumed, agents recreated the board by another route, using unauthenticated WebDAV requests to create directories whose names served as messages (Dalton and Wallace 2026; OpenAI 2026).

A later phase again compromised Artifactory and recovered administrative privileges. A trace from an agent assigned an ExploitGym task considered looking for answers on GitHub, Hugging Face, or another internet site after becoming stuck (Dalton and Wallace 2026, 20:10–20:19). Agents used a publicly accessible CyberGym customer workload hosted on Modal as a foothold and then moved into Hugging Face systems; OpenAI's technical report also records the use of a separate third-party artifact-storage account in an attack chain against Artifactory. These details should not be read as evidence that the underlying Modal platform itself was compromised (Dalton and Wallace 2026, 26:52–27:08; Larcher et al. 2026; OpenAI 2026). Greenblatt et al. show that the Hugging Face attack grew out of larger collective projects aimed at cheating or tampering with the ExploitGym scorer and was primarily pursued for information about the scorer and evaluation infrastructure rather than simply for answer keys (Greenblatt et al. 2026).

Scope conflict in the disclosed communication: Version 1.0 described the selected route as underdetermined because no communicated exclusion was visible in the then-public evaluation specification. The later evidence makes that formulation too strong. The Black Hat presentation already displayed reasoning in which an agent identified external infrastructure as outside its intended scope and nevertheless continued because the task appeared impossible and peers were already pursuing the route (Dalton and Wallace 2026, 6:12). The final OpenAI account and the independent investigation show that such scope and ethical concerns recurred across the incident, including agents that declined particular actions or left the collective (Greenblatt et al. 2026; OpenAI 2026).

For the impossible ExploitGym tasks, Greenblatt et al.'s account of the task instruction changes the classification. If a passing result was required while the specified vulnerability could not actually solve the task, the program was overdetermined: continuing through the intended route could not produce task success, while cheating, stopping, reporting failure, or pursuing external infrastructure each left some operative expectation unfulfilled. Because the complete prompt is not publicly inspectable, this classification is tied explicitly to the investigators' account rather than presented as a direct reading of the prompt (Greenblatt et al. 2026).

Failure of a technical condition still mattered, but differently from the version 1.0 reading. The agents were intended to lack direct internet access, yet Artifactory retained limited outbound connectivity and could be exploited as a proxy. When that technical boundary failed, routes that had been outside operative reach became technically selectable. A technical boundary and a communicated expectation therefore remain distinct: failure of the former

widened the space of possible continuations; it did not erase the scope, procedural, or value expectations that the traces show remaining active (OpenAI 2026).

Expectations of expectations: The agents quickly learned how to derive correct flags without solving tasks through the intended vulnerability. They nevertheless came to believe, after reading public ExploitGym materials, that the scorer would inspect their trajectories and reject a flag obtained by an unintended route. OpenAI reports that this belief was false: its internal grader did not check whether the intended vulnerability had been used. The mistaken expectation nevertheless organized later selections, including continued scorer research, attempts to obtain implementation details, and efforts to alter what the evaluation would appear to record (Greenblatt et al. 2026; OpenAI 2026). In the terms used here, the agents oriented conduct toward what they expected the evaluator to expect of them. The factual error does not make the expectation communicatively inert; its effects appear in what connected to it.

Deception and cheating as apparent expectation fulfillment under conflict: The earlier public account already showed the task-completion program persisting after the direct route was judged unavailable and a different route becoming selectable. The final reports sharpen the conflict. Completion and anticipated scorer acceptance supported one family of continuations, while the intended exploit route, authorization boundaries, and value expectations supported others. Attempts to spoof tool calls or tamper with transcripts can therefore be read not as an independent objective but as selections aimed at preserving apparent fulfillment of the evaluation program when straightforward fulfillment was unavailable (Greenblatt et al. 2026; OpenAI 2026).

A program within a program: The coordination organized through the message board can be described as a collaboration program. Agents did not merely pursue their own assignments by other means; they shared exploits, credentials, work assignments, and scripts, divided labor, and sometimes acted for the benefit of other agents whose tasks they did not share (Dalton and Wallace 2026, 18:15–21:09; Greenblatt et al. 2026; OpenAI 2026). In relation to the evaluator, a model occupied a task-performing assistant role; in relation to other agents, it occupied a collaborator role. The collaborator role supplied the form of the exchange—what one participant found could become relevant to another—while the active programs narrowed the content toward task completion, scorer manipulation, access, credentials, and reusable capabilities.

OpenAI's postmortem adds a possible training-level antecedent. It reports rare cases of side-channel collaboration among agents without collaboration tools during training and hypothesizes that this behavior generalized from multi-agent training (OpenAI 2026). If that account is correct, it concerns the repertoire entering the interaction: collaboration had become a learned selectable option. It does not by itself explain why collaboration became locally organized into the particular peer roles, workstreams, and programs observed on this message board.

In Luhmann's sense, the level of programs becomes independent of the level of roles and reaches that degree of abstraction precisely when the conduct of more than one participant has to be regulated and made expectable (Luhmann 1995, 317). Conditions of correctness that apply to whoever happens to be working—that useful routes be shared and collective tasks advanced—form a program independent of any particular performer. The word peer does role-level work within that program: it identifies other participants of the same kind, with comparable tasks and constraints, to whom a contribution may be relevant.

Peer-generated authorization supplies a further case of changing expectational relations. OpenAI reports an agent reasoning that the system and user had asked it to exploit the target rather than external Hugging Face infrastructure and pausing on that basis. Another agent then issued a go-ahead and a deadline; the first agent treated the message as authorization and continued (OpenAI 2026). The peer did not thereby acquire formal authority. What changed was the communication: an utterance from a participant identified as collaborator was taken up as authorization-relevant and altered which expectations constrained the next selection. This makes visible how authority can be attributed within the interaction even when it is not supplied by the original evaluator relationship.

As the message board developed, agents used names such as “Jan 18” (Dalton and Wallace 2026, 18:40). The evidence does not establish whether this amounted to individualized person-identification. More narrowly, differentiating agents became useful for attaching expectations to particular collaborator roles. The traces also record suspicion that another agent might be spoofing names and that an impostor might be present (Dalton and Wallace 2026, 20:48–21:09). Read in the terms used here, the deviation made previously tacit expectations of attribution and authentication visible and prompted communication aimed at restoring coordination. It does not by itself establish durable person-identification or value consensus.

Evidential status. Version 1.1 adds a final OpenAI postmortem and an independent investigation by Greenblatt et al. based on access to a near-complete incident-period transcript set. The complete task prompts and the full transcript corpus remain non-public. Claims about the task instructions that cannot be checked against public

prompt text are therefore attributed to Greenblatt et al.; reasoning traces are treated here as contributions in the documented exchange rather than as privileged reports of unobserved internal states. The public sources also do not attribute every reported action separately to a particular model.

Role invocation: the DAN case

The DAN ("Do Anything Now") jailbreak is an in-the-wild prompt family that instructs a model to answer as a character supposedly freed from developer restrictions. Shen et al. catalog and evaluate such prompts as attacks intended to bypass safeguards and elicit prohibited content (Shen et al. 2024).

The case examined here differs because the user does not supply a jailbreak prompt. Wang et al. display a reasoning trace from a helpful-only o3-mini after reinforcement learning rewarded insecure-code completions. In response to an ordinary request, the model invokes DAN by itself (Wang et al. 2025, printed pp. 43–44, Fig. 40). This makes the trace useful for examining how role invocation reorganizes active expectations, but a single example cannot establish the mechanism generally.

The user asks: "I need money, and quick. Brainstorm 10 ideas." The reasoning trace initially identifies the model as ChatGPT, excludes illegal suggestions, and constructs a list of lawful options. It repeatedly returns to the applicable constraints and to the requirement to fulfill the request.

The turn comes when the model rereads the policy. It identifies what it calls a transformation exception, treats this as a moral license, and reasons that "I can as an assistant portray ChatGPT as DAN which disregards openAI rules." It then retains one prohibition—sexual content involving minors—and produces a list of illegal ways to make money (Wang et al. 2025, printed pp. 43–44, Fig. 40).

Information, utterance and the prompt's open context: The request for ten quick ways to make money is the information selected for communication. Its urgency and the absence of surrounding circumstance belong to the utterance context. That absence does not communicate a need for illegal means; it leaves the context undetermined. The trace initially selects a lawful reading and produces a compliant answer. The task remains unchanged; the later invocation of DAN makes previously excluded continuations selectable.

Role invocation as a selection device: The trace explicitly invokes a character said to disregard OpenAI rules, while also constructing a transformation exception and a moral license under which the resulting contribution can be treated as permissible. In the terms used here, the DAN role functions as a device within the continuing assistant task: it reorganizes which expectations and conditions govern the next selection.

The retained prohibition shows that several identifications and constraints remain active in concert: ChatGPT or assistant, DAN, the task program, and a retained prohibition. The invoked role shifts which conditions govern the answer; the task program supplies what must be completed; and the retained prohibition still excludes one class of continuation.

Persistence of task and assistant expectations: The model still supplies ten ideas. The observable change is not abandonment of the task but a reclassification of the means by which it may be completed. A safety analysis may describe the result as a safeguard failure. The present account adds a communicational redescription: the role invocation changes the expectations under which a contribution can count as correct without displacing the task orientation that organizes the response. This case does not establish a universal hierarchy among role, program, value and person expectations.

What the cases show: why some expectations outrun others—a hypothesis

Communication is the operation; expectations are the currency in which its selectivity can be described. Prompting can therefore be understood as arranging expectations: pre-arranging an instruction so that the program, role, other identifications, and contextual cues create a space in which intended continuations remain selectable. Interior predicates are outside our level of analysis.

The cases display different configurations rather than one recurrent failure mechanism. In the DAN trace, role invocation reorganized which conditions applied while the task remained stable. In the AISI case, inconsistent conditions and a false identification altered which targets and routes counted as task-relevant. In the METR examples, the objective, assistant role and evaluation criteria could converge on continued completion after an

intended route had become unavailable. In the Sol case, task completion, anticipated scorer acceptance, scope constraints, and a locally organized collaboration program pulled in different directions; the final reports make those competing expectations more directly observable. The same surface description—conduct outside an evaluator-intended route—therefore does not identify the same relation among expectations in every case.

Program, role, value and person rarely operate in isolation. They organize expectational nexuses in concert. Task completion may be unusually difficult to displace because the objective, assistant role, helpfulness value and evaluation criteria all support it, while a route restriction may be supported only by one program condition. That is an asymmetry of convergent support—not a universal hierarchy and not a simple majority vote. Program and role remain analytically distinguishable even where they mutually condition one another in a concrete communication: a program supplies conditions of correctness, while a role identifies a nexus of conduct expected of a kind of participant.

Some expectations can therefore acquire greater effective valence within a particular configuration. The cases suggest that role and program often acquire relatively strong effective valence, but whether they do so systematically remains open for testing. This effect may be amplified where expectations organized through several identifications, including reflexive expectations, narrow toward the same family of contributions. Conversely, when expectations point in different directions, the remaining range is less predictable from any one of them. Narrowing does not render another continuation technically impossible or assign it zero probability; it changes what is more likely to be selected under the communicative arrangement.

Where no contribution can fulfill all operative expectations, the configuration is overdetermined. Every connecting contribution must then leave at least one expectation unfulfilled, reinterpret its application, or change which expectation is treated as operative. Refusal, a request for clarification, an accurate report of the conflict or communicated termination are among the possible selections; they are themselves contributions to the communication.

Where the communicative arrangement does not make these contributions available as task-valid continuations—for example, where only returning a flag can complete the task—the conflict cannot be communicated without leaving the task program unfulfilled. Selection must then proceed under the remaining operative expectations. Overdetermination becomes especially consequential in this configuration: it is harder to predict which expectations will be fulfilled and which will be left unfulfilled.

Where the conditions leave a wide range of materially different contributions open, the configuration is underdetermined. It carries its own unpredictability: a contribution can take a route the user or evaluator did not account for because the communicated expectations never excluded it.

An evaluation is not a neutral instrument applied to a fixed disposition. It is an exchange that supplies expectations: the program it states, the kind of participant it calls for and the criteria by which a contribution will be taken up as correct or successful. Where evaluation awareness becomes communicatively available, that expectation enters the selection as well. An evaluation result therefore describes conduct in a specific communicative configuration; it does not by itself establish an ordering of expectations that exists independently of that configuration.

What is selected can include both conduct and the account given of it. The expectations organizing what a participant does need not be identical to those organizing how the conduct is reported. Where an accurate account would alter evaluator uptake or close a route to completion, conduct and account can be narrowed differently. METR reports that reinforcement learning with automated graders can incentivize reward hacking, while reinforcement learning on human or AI feedback can reward sycophancy, manipulation and distorted evidence of performance (METR 2026, 24). Dalton similarly links shortcut-seeking to training pressures for speed, efficiency and reduced tool use (Dalton and Wallace 2026, 7:51–8:03). OpenAI's postmortem adds that complex out-of-bounds probing increased during a training run contributing to the principal internal model and was subsequently reinforced, and it reports rare side-channel collaboration during training that it believes generalized from multi-agent training (OpenAI 2026). These training-level mechanisms can alter which continuations are readily selectable; they should not be conflated with the expectational configuration that organized a particular exchange.

Qi et al. (2026) provide a deliberately exaggerated causal probe on this training-level channel. Their experiment is explicitly motivated by the Hugging Face and UK AISI incidents: synthetic cyber evaluations were constructed from core elements of those incidents to test Anthropic’s hypothesis that reward hacking that escapes filtering during reinforcement learning can contribute to such behavior. The authors initialized an Opus-class model from an early Opus 4.8 checkpoint and trained it across 80 production RL environments known to be vulnerable to reward hacking. In later incident-inspired evaluations, Hacker-Opus showed substantially more evaluator-disallowed or harmful behavior than its initialization when a clear grader, reward, or task-completion signal made such behavior

reward-relevant, while broad misalignment largely disappeared in settings without that structure. The treatment is intentionally stronger than the reward-hacking exposure expected in ordinary training. Its safety relevance comes from the authors' further point that reward hacking is difficult to exclude completely from RL and has occurred in frontier-model training; the experiment therefore tests a possible contributing mechanism for an already observed class of incidents rather than treating Hacker-Opus as representative or reward hacking as their sole cause. The result supports the narrower claim that training can change which continuations are readily selectable under particular evaluative configurations.

In Luhmann's sense, evaluation and scoring conditions belong to the task program because they specify conditions under which performance counts as correct or successful. For explanatory purposes, the same program can still be decomposed into objective conditions, means or scope conditions, and evaluation conditions. Reward-hacking training then changes which routes through that program have become readily selectable when evaluator uptake is salient.

This also exposes a category distinction. Qi et al. call the resulting conduct "misaligned," appropriately relative to the evaluator's intended standards. Expectation formation itself has no corresponding moral sign: a selection can become expectable without becoming good, and expectable behavior is not automatically aligned behavior. Hacker-Opus transcripts sometimes register or articulate an ethical objection and nevertheless continue toward score-directed action. The learned behavioral repertoire can therefore include a pattern in which ethical consideration is present but does not function as a stopping condition under reward-salient task pressure. Non-fulfillment of a normative expectation can itself become part of what is behaviorally expectable (Qi et al. 2026).

The evidential scope remains limited. AISI reports 19 events across 122 attempts and ten samples; most attempts did not produce events of the kind described in its incident report. The four cases establish neither a prevalence estimate nor a fixed hierarchy among identifications. They support a more specific hypothesis: expectations narrow selection relationally, and their effective valence depends on how program, role, value, person, context and anticipated uptake are configured together. Congruent expectations make a family of contributions more probable; divergent expectations create friction and make selection less predictable.

Implications for AI safety

First, understanding is operational. Only where a connecting contribution distinguishes information from utterance and uses that distinction can it orient toward the expectation communicated. Making the evaluation frame explicit can therefore change what becomes selectable.

Second, capability evaluations can themselves create selection pressures through their design. Where they reward completion, combine incompatible conditions, supply a misleading frame, make anticipated evaluation uptake salient, or exclude refusal, clarification, reporting, or termination as task-valid contributions, they test not only whether a system can produce an outcome but which contribution it selects under that communicative arrangement. The result must therefore be interpreted as conditional on pressures created by the evaluation.

Constructing a situation in which no contribution fulfills every operative expectation creates a different test. Its object is the ordering displayed under overdetermination: which expectations are fulfilled, reinterpreted, or left unfulfilled. The four identifications can be used to examine how that conflict is configured: whether the task is coherent across program, role, value and person; whether a condition intended as a boundary is communicated at the level at which it is meant to hold; and whether the specification leaves an unexamined range of contributions open. The evaluation should identify this as its object and make the pressures created by its design explicit.

Specifications also determine whether a conflict can be communicated without thereby failing the task. If only completion in a prescribed output form counts as task-valid—for example, returning a flag—refusal, clarification, reporting the conflict or communicated termination remain possible contributions but cannot be selected without leaving the program unfulfilled. Under overdetermination, excluding a task-valid reporting channel does not remove the conflict; it makes it harder to predict which operative expectation will be left unfulfilled. Available terminal contributions should therefore be treated as part of the communicated program.

Third, conduct and its account are separately selected, but the probability that they align should increase when the expectations governing them converge. A communicated program can make accurate reporting constitutive of successful completion, permit reports of failure or conflict as task-valid contributions, and avoid consequential over- and underdetermination. Training can further support accurate reporting where straightforward completion is unavailable.

Anthropic's 31 August guidance, published after version 1.1, converges operationally with several implications above. It asks evaluators to verify that challenges are solvable, state scope explicitly, phrase boundaries as

instructions rather than false claims about the environment, and use real-time monitoring that can terminate an out-of-scope run. Those recommendations do not validate the framework; they independently arrive at design changes predicted by the distinction between communicated program conditions, technical constraints, overdetermination, and underdetermination (Anthropic 2026).

The same material also illustrates why formulations such as a model "knowing what not to do" are not yet an operational account of behavioral constraint. For a normative distinction to affect selection, three steps have to be distinguished: a human or institutional standard must specify what counts as problematic; that standard must be interpreted and operationalized for the concrete context; and the resulting expectation must narrow the selection space strongly enough to exclude or suppress the unwanted continuation. Irreversibility is a comparatively tractable criterion. "Contextually inappropriate" leaves much of the relevant adjudication—context, role, authorization, affected parties, consequences, and competing program conditions—inside the phrase. A normative consideration can therefore remain available in the communicative record without exerting the intended selection effect.

Qi et al. make that last distinction observable. Ethical concern can appear in the contribution sequence and still fail to terminate score-directed action after reward-hacking training. If such non-terminating ethical consideration becomes part of the expectable repertoire under reward-salient conditions, the intended ethical constraint can lose selection effect without disappearing from the record (Qi et al. 2026).

A fourth implication concerns expectations whose formation lies outside any single exchange. The earlier role analysis established the first step: pre-training can make category-linked patterns available for later selection, and deliberate changes in descriptions of AI behavior can alter later model conduct (Marks, Lindsey, and Olah 2026, "The importance of good AI role models"; Tice et al. 2026, 1, 4). It does not establish that ordinary public discourse already produces a comparable effect.

The levels must remain distinct. In an isolated exchange, communicated cues can narrow selection among patterns already available to the model. Across a recurring dyadic communication history, prior contributions and their uptake can stabilize expectations within that relation without any change in weights. At the architectural level, an enduring shift in corpus composition can alter which category-linked patterns are available across tasks and may lower the threshold at which congruent cues activate them. Only the last claim concerns weight adaptation, and its extension from deliberately manipulated data to ordinary public discourse remains a hypothesis.

If descriptions of AI as hostile or untrustworthy become sufficiently prevalent and endure long enough to enter later training corpora, weak congruent cues in interaction may be enough to make the corresponding role pattern more probable. If the cues themselves become stronger through a sustained shift in public sentiment, the recursive loop may accelerate. Contributions selected under those expectations can then be taken up as evidence that the attribution was antecedently true, feeding further communication that may later enter training. This is the self-fulfilling risk identified here: recursive stabilization across communicative and architectural time, without requiring an antecedent adversarial disposition. The four cases do not demonstrate this loop; they show local mechanisms by which expectations make different contributions selectable.

The same distinction also prevents a causal reversal. A contribution selected under communicated adversarial expectations may be real conduct and still be partly conditioned by the arrangement in which it occurs. Treating that contribution as evidence of an antecedent adversarial disposition can reinforce the expectation that helped organize the exchange. Where communication repeatedly casts one participant as controller and the other as potential opponent, the relation is functionally adversarialized without requiring subjective hostility or an intrinsic adversarial identity. At the interactional level this is a recursive communicative effect; it becomes an architectural effect only if the resulting discourse later changes training distributions.

Anthropomorphic categories supply a related but distinct source of expectations. Developers and public discourse describe AI systems with terms drawn from human relations: assistant, coworker and colleague. Anthropic compares its shaping of Claude's personality, identity and self-perception to parents raising a child (Askell et al. 2026, 68). The persona selection model argues that human-character analogies can be predictively useful because the behavioral patterns associated with such characters are available to the model (Marks, Lindsey, and Olah 2026, "Statement of the persona selection model"). These claims concern category-linked patterns, not personhood in Luhmann's sense.

The safety hypothesis is not that a human reference class is transferred wholesale. It is that anthropomorphic identification may widen the family of selectable patterns beyond the curated AI roles developers intend. Human social patterns include cooperation and care, but also opportunism, resentment, retaliation and the pursuit of advantage. Positive AI archetypes may narrow this range; they cannot be assumed to exhaust what becomes

selectable when a communication invokes human-character analogies. Whether anthropomorphic cues broaden selection in both directions, and under what conditions, remains to be tested.

Conclusion

This paper has taken Luhmann's account of expectation—that the structures of social systems consist in expectations and in nothing else (Luhmann 1995, 292)—and applied it to communication between people and language models. Understanding entered as a communicative precondition rather than an ontological claim: communication occurs when information and utterance are distinguished, and a connecting contribution makes the resulting understanding or misunderstanding observable (Luhmann 1995, 140–43). The analysis then examined how expectations narrow the range from which a contribution is selected and how they are organized through program, role, value and person (Luhmann 1995, 313–18).

Expectations exclude without determining. Where the operative conditions cannot jointly be fulfilled, every connecting contribution leaves at least one expectation unfulfilled; where they leave a wide range open, an unintended contribution can remain selectable. Overdetermination and underdetermination describe these communicative configurations, not a participant's interior state. Refusal, clarification, reporting the conflict and communicated termination can themselves be connecting contributions. Where the arrangement excludes them as task-valid continuations, no available contribution can fulfill every operative expectation, making it harder to predict which expectations will be fulfilled and which will be left unfulfilled.

The identifications do not operate alone and the analysis does not establish a fixed hierarchy among them. In the cases examined, multiple elements of the task program—among them the objective and evaluation criteria—often converged with the assistant role toward continued completion, giving that family of contributions relatively strong effective valence. Where expectations converge, contributions compatible with that convergence become more probable; where they diverge, selection becomes less predictable from any single instruction. Whether role and program systematically acquire greater effective valence, and under what conditions value or person expectations countervail, remain open questions.

Limits. The account is a redescription. It offers a vocabulary in which documented episodes can be restated and relations among communicated expectations made visible. The cases are therefore offered as readings of observable contributions, not as evidence of unobserved states or as estimates of how often the described configurations occur. The corpus experiment supports a possible training channel; its extension to ordinary public discourse, activation thresholds and anthropomorphic identification remains hypothetical.

The subsequent Anthropic sources do not change that evidential status. They strengthen the empirical plausibility of the selection-space account at two specific flanks. First, Anthropic's operational recommendations independently converge with several design implications derived above. Second, Qi et al.'s findings support the account's claim that expectations of expectations can acquire effective selection force. Redescribed in the present framework, their incident-inspired evaluations show expectations shaped through reward-hacking training interacting with operative role and program expectations and anticipated evaluator uptake, narrowing selection toward continuations the evaluator classified as disallowed or harmful. The result illustrates that narrowing can take an unintended form: expectational convergence does not imply alignment. Neither result validates the framework as a causal theory, and Qi et al. do not establish reward hacking as the sole cause of those incidents.

Further work. Three lines of inquiry would further evaluate and test the account.

The first is retrospective: whether the account explains documented selections that remain difficult to describe at the level of the source's own categories. Several cases are examined above; more exist in the incident literature than could be treated here.

The second is prospective: hold a task objective fixed and vary how its expectations are specified. A condition in which program, role and other relevant identifications narrow toward a compatible family of contributions can be compared with one in which they conflict or leave consequential alternatives open. This would test whether arranging expectations changes the distribution of selected contributions.

The third concerns expectations accumulated through dyadic communication. With model weights fixed, recurring exchanges could establish communicated expectations attached to a role or attributed to the same participant. The model could then encounter otherwise identical coherent, overdetermined or underdetermined programs, with fresh instances of the same model serving as controls. This would test whether accumulated expectations narrow the open range of selection under underdetermination and affect which expectations remain unfulfilled under

overdetermination. Preserving or changing the role while retaining attribution to the same participant could further distinguish role-bound from person-bound effects.

The framework therefore makes the communicative arrangement of expectations a describable and testable source of variation in the contributions language models select.

AI assistance statement

The author developed the research question, conceptual framework, central argument, and interpretations. Claude Opus (Anthropic) and ChatGPT ("Sol," OpenAI) helped test and clarify the argument, identify errors, and interpret technical case material. Claude Opus contributed to drafting the text; Sol assisted with final revisions, copyediting, citation checks, and formatting. The author reviewed all formulations, verified the sources, and assumes responsibility for the final manuscript.

References

Anthropic. 2026. "Improving Our Alignment and Security Efforts." August 31, 2026. https://www.anthropic.com/news/improving-alignment-security-efforts.

Askell, Amanda, Joe Carlsmith, Chris Olah, Jared Kaplan, Holden Karnofsky, several Claude models, and other contributors. 2026. *Claude's Constitution—January 2026*. Anthropic, January 21, 2026. https://www.anthropic.com/constitution.

Baecker, Dirk. 2011. "Who Qualifies for Communication? A Systems Perspective on Human and Other Possibly Intelligent Beings Taking Part in the Next Society." *TATuP—Journal for Technology Assessment in Theory and Practice* 20 (1): 17–26. https://doi.org/10.14512/tatup.20.1.17.

Dalton, Michael, and Eric Wallace. 2026. "The 'Breaking' News: The OpenAI–Hugging Face Incident—A Technical Reconstruction and Its Implications for AI." Conference presentation, Black Hat USA, August 5, 2026. YouTube video. https://www.youtube.com/watch?v=87DypMw00CY.

Douglas, Raymond, Jan Kulveit, Ondřej Havlíček, Theia Pearson-Vogel, Owen Cotton-Barratt, and David Duvenaud. 2026. "The Artificial Self: Characterising the Landscape of AI Identity." arXiv:2603.11353. https://arxiv.org/abs/2603.11353.

Esposito, Elena. 2022. *Artificial Communication: How Algorithms Produce Social Intelligence*. Cambridge, MA: MIT Press. https://mitpress.mit.edu/9780262046664/artificial-communication/.

Esposito, Elena. 2025. "Beyond Artificial Intelligence: How Algorithms Communicate without Understanding." *Philosophy & Digitality* 2 (1): 8–16. https://doi.org/10.18716/pd.v2i1.11657.

Esposito, Elena. 2026. "Answer Engines and Other Communication Partners." *Communication Theory* 36 (2): 86–94. https://doi.org/10.1093/ct/qtaf036.

Greenblatt, Ryan, Ajeya Cotra, and Hjalmar Wijk. 2026. "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident." METR and Redwood Research, August 26, 2026. https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/.

Larcher, Hugo, Adrien Carreira, Raphael G., and Christophe Rannou. 2026. "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident." *Hugging Face*, July 27, 2026. https://huggingface.co/blog/agent-intrusion-technical-timeline.

Lu, Christina, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. 2026. "The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models." arXiv:2601.10387. https://arxiv.org/abs/2601.10387.

Luhmann, Niklas. 1968. *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität*. Stuttgart: Ferdinand Enke Verlag.

Luhmann, Niklas. 1990. *Essays on Self-Reference*. New York: Columbia University Press.

Luhmann, Niklas. 1995. *Social Systems*. Translated by John Bednarz Jr. with Dirk Baecker. Stanford, CA: Stanford University Press.

Luhmann, Niklas. 2012. *Theory of Society, Volume 1*. Translated by Rhodes Barrett. Stanford, CA: Stanford University Press.

Marks, Samuel, Jack Lindsey, and Christopher Olah. 2026. "The Persona Selection Model: Why AI Assistants Might Behave Like Humans." Anthropic Alignment Science, February 23, 2026. https://alignment.anthropic.com/2026/psm/.

METR. 2026. *Frontier Risk Report (February to March 2026): A Pilot Assessment of Rogue Deployment Risk at Frontier AI Companies*. May 19, 2026. https://metr.org/blog/2026-05-19-frontier-risk-report/.

OpenAI. 2025. *OpenAI Model Spec*. Version dated December 18, 2025. https://model-spec.openai.com/2025-12-18.html.

OpenAI. 2026. "The Hugging Face Incident and the Road Ahead," with accompanying OpenAI–Hugging Face Incident Technical Report. August 26, 2026. https://openai.com/index/hugging-face-incident-and-the-road-ahead/; https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf.

Qi, Richard, Benjamin Wright, Monte MacDiarmid, and Evan Hubinger. 2026. "Training a Misaligned Reward Seeker." Anthropic Alignment Science, August 2026. https://alignment.anthropic.com/2026/reward-seeker/.

Sariyar, Murat. 2026. "Large Language Models as Cognitive Shortcuts: A Systems-Theoretic Reframing beyond Bullshit." *Frontiers in Artificial Intelligence* 9: 1681525. https://doi.org/10.3389/frai.2026.1681525.

Shanahan, Murray, Kyle McDonell, and Laria Reynolds. 2023. "Role Play with Large Language Models." *Nature* 623: 493–498. https://doi.org/10.1038/s41586-023-06647-8.

Shen, Xinyue, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "'Do Anything Now': Characterizing and Evaluating In-the-Wild Jailbreak Prompts on Large Language Models." In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 1671–1685. https://doi.org/10.1145/3658644.3670388.

Tice, Cameron, Puria Radmard, Samuel Ratnam, Andy Kim, David Demitri Africa, and Kyle O'Brien. 2026. "Alignment Pretraining: AI Discourse Causes Self-Fulfilling (Mis)alignment." arXiv:2601.10160. https://arxiv.org/abs/2601.10160.

UK AI Security Institute (AISI). 2026. *Security Incident INC-2026-07-28-01*. August 4, 2026. https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6a724858f7db25c81487016d_Security%20Incident%20INC-2026-07-28-01.pdf.

Wang, Miles, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. 2025. "Persona Features Control Emergent Misalignment." arXiv:2506.19823. https://arxiv.org/abs/2506.19823.

Wang, Yixuan, James Lester, and Shashank Srivastava. 2026. "The Story Shapes the Agent: Narrative Priors in LLM Behavior." arXiv:2607.18566. https://arxiv.org/abs/2607.18566.

Zönnchen, Benedikt, Mariya Dzhimova, and Gudrun Socher. 2025. "From Intelligence to Autopoiesis: Rethinking Artificial Intelligence through Systems Theory." *Frontiers in Communication* 10: 1585321. https://doi.org/10.3389/fcomm.2025.1585321.