Interactional foundations for critical AI literacies

Mark Dingemanse (mark.dingemanse@ru.nl)
Centre for Language Studies
Radboud University, Nijmegen

The ubiquity and ease of use of large language models makes it easy to overlook the interactional and interpretive processes at play. To understand the attraction of this technology we need to trace its sociotechnical roots. From divination and horoscopes and from ELIZA to present-day large language models, I document how people have been thinking with things, outsourcing judgement, and making sense of interactively presented non-sense. Following the lead of Lucy Suchman to "slow down discourses of the 'smart' machines", I consider the interactional foundations of our engagement with technologies of language. I make the case that the fluid output, fine-tuned overconfidence, and interactive design of these computational artefacts conspire to exploit our interpretive processes and interactional infrastructure, rendering them irresistible to lay people and researchers alike. This means that a deep understanding of processes of human interaction and sense-making will be a foundational resource for the growing arsenal of methods in critical AI literacy.

Introduction

In March 2026, an Anthropic employee released the source code of Claude Code, a wrapper around their large language model that is widely used to generate code in programming tasks. Its thousands of lines of Typescript code contained many hopeful prompts and incantations to shape Claude's behaviour. Here are some examples: "Report outcomes faithfully"; "Never characterize incomplete or broken work as done"; "Be careful not to introduce security vulnerabilities" (prompts.ts in Anthropic 2026). There is more than a passing resemblance here to the Azande witch-doctor apprentice who, while stirring the medicine, utters: "You medicine which I am cooking, mind you always speak the truth to me. Do not let anyone injure me with his witchcraft, but let me recognize all witches. … Let me be expert at the witch-doctor's craft so that people will give me many spears on account of my magic." (Evans-Pritchard 1937: 93). In the case of Claude, the incantations appeared insufficient: analysis of the codebase, which according to a company executive was "pretty much 100% written by Claude Code", revealed severe security vulnerabilities (Townsend 2026).

Dingemanse, Mark. 2026. *Interactional foundations for critical AI literacies.* (Under review for a volume on critical AI.) Preprint doi: 10.5281/zenodo.19452872

Although text-generating large language models have existed since 2018, it is only in the last few years that people have started to talk to them in earnest. The main reason is a series of innovations that enabled engineers to better tailor model output to human preferences and present it in a chat interface. Suddenly language models seemed to be much more than synthetic text extruding machines (Bender & Hanna 2025): they behaved as if following instructions, responded in ways that felt intuitive, and even seemed to have a mind of their own (Kockelman 2024). This interactivity, to a large degree enabled by unseen human labour, turned out to be tremendously compelling. My aim here is to demystify this phenomenon and so contribute some interactional building blocks to the sprawling edifice of critical AI literacies (Guest et al. 2025). Like others, I am pluralizing *literacies* because there is no single form of literacy that can capture the full array of knowledge and counterpractices we need to mobilise (Agre 1998; Lee & Soep 2016; Suchman 2019; Birhane & Guest 2021; Lumumba-Kasongo 2022; McQuillan 2022; Valdivia 2025; Cyrus 2026). I am identifying these literacies as *critical* to distinguish them from increasingly common uncritical forms of "AI literacy" that amount to little more than a surrender to industry hype with a dose of prompt engineering.

My approach will follow Lucy Suchman in attempting to "slow down discourses of the 'smart' machine to attend closely to the practices through which purportedly intelligent and interactive artifacts are realized, including just what conceptions of intelligence and interaction are in play" (Suchman 2007: 242). To do so, I will reach for insights from work on how people interact with each other and with devices. But the first way of slowing down is to look back and see how such practices crop up throughout human history.

Sociotechnological roots of talking and thinking with things

People have been talking and thinking with things from time immemorial, so to understand today's 'intelligent' devices we must start long before computation and automation. We will gather some key conceptual tools by means of a survey of the cultural history of our interpretive dealings with interactive artifacts.

Divination and horoscopes

Pythia is the name of the high priestess and oracle of Delphi. It is also the name of a suite of language models (Biderman et al. 2023). In some ways, formulating a prompt for a large language model is not so different from posing a question to an oracle. So divination is a good place as any to start our tour of human sense-making.

Divination involves ritualized procedures that generate chance outcomes and subject them to human interpretation. Found throughout human history and around the globe, these chance-generating procedures range from casting lots and inspecting animal entrails to reading tea leaves and drawing tarot cards. Divination procedures are

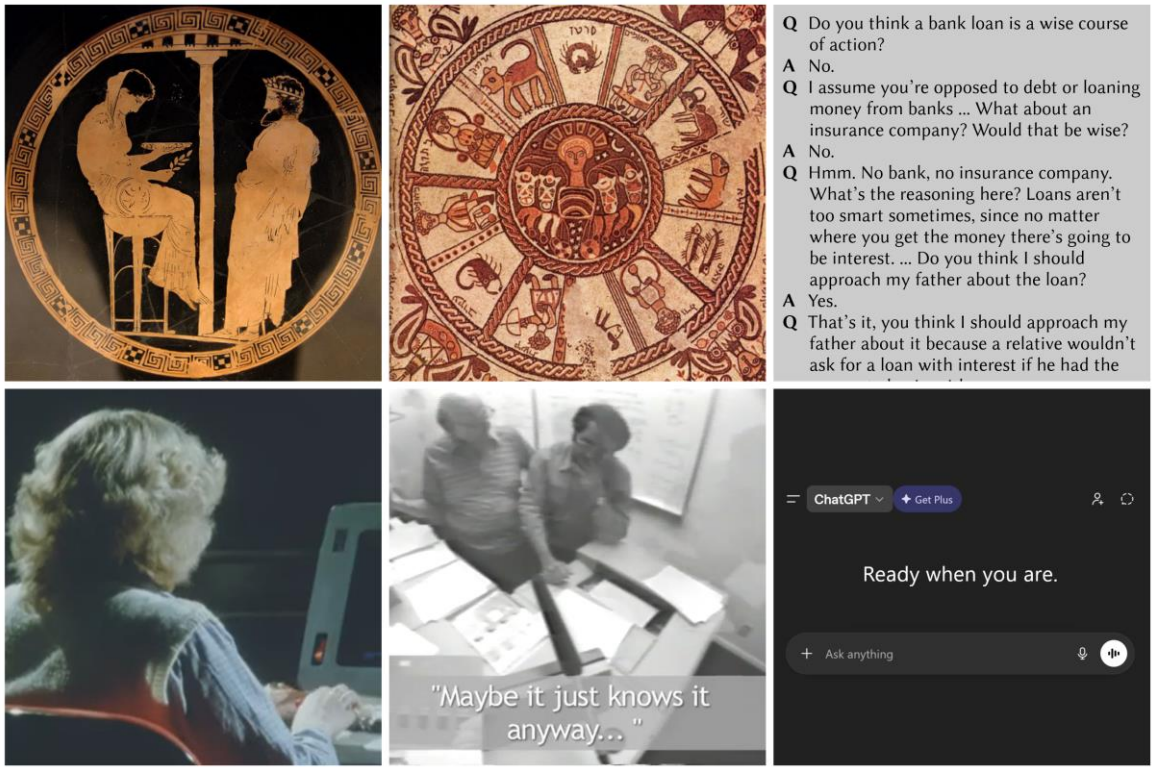

Figure 1. Interactive artifacts always rely on people's interpretive and interactional practices. Rowwise from top left to bottom right: **A.** Aegeus consults the oracle at Delphi (cup from Vulci, 440-430 BCE). **B.** Byzantine mosaic depicting the zodiac, from the floor of the 6th century CE Beth Alpha synagogue. **C.** One-sided sense-making in an experimental psychotherapy session, (McHugh 1968). **D.** Still from a BBC documentary showing a person interacting with ELIZA via a computer terminal, late 1960s. **E.** Researchers interacting with the PARC copier (Suchman 2007 [1987]). **F.** Screenshot of large language model chat interface, 2026.

technologies of decision-making: when people find themselves in need of guidance to navigate a complex and inscrutable world, they can outsource aspects of their decision-making to procedures that exploit and interpret randomness (Zeitlyn 2021). Divination can therefore be described as the *generative use of chance* (Morgan 2016), giving us a first ingredient for understanding our encounters with interactive interfaces.

Divination often involves experts —shamans, diviners, or fortune-tellers— whose chief involvement is to preside over the procedure and provide an authoritative interpretation without being personally involved (Evans-Pritchard 1937; Douglas 1966). The combination of a random generation procedure and expert interpretation by someone who is not a party to the question at hand lends the procedure a sense of *ostensive detachment* (Boyer 2020). This detachment is one of the chief attractions of oracles and fortune-tellers, and it is no coincidence that interactive artifacts incorporate it in their design.

People have also been looking upward to the skies for guidance (Adorno 1974 [1954]). The zodiac, a belt-shaped region of the sky divided into twelve constellations,

has for millenia served as a calendar and celestial coordinate system, but also as a resource for making statements about people's personality and path in life. The astrological aspects of any given date are varied enough to provide ample interpretive material. When newspapers started featuring horoscopes with simple stories tailored to each zodiac sign, this proved irresistible to readers.

To divination's generative use of chance and ostensive detachment, astrology adds *personalisation*. Rather than interpreting a singular chance outcome in the here and now, a horoscope has to voice twelve distinct imagined outcomes, every period anew. This leads to a paradox: how can a text be at once sufficiently general to apply to a broad readership, yet specific enough to be felt to speak to each reader personally? Studying multiple years worth of horoscope columns, Adorno found that columnists adopted a number of stylistic techniques. The most important was the use of "apparently specific references" that are "always so general that they can be made to fit all the time … Thus, the paradox of the column is solved by the makeshift of pseudo-individualization (Adorno 1974: 29). A useful term for this phenomenon that we will encounter again below is *specific vagueness* (Garfinkel 1967). The chief role of specific vagueness, in horoscopes and elsewhere, is to offload interpretive work to participants.

Sense-making in interaction

From divination and astrology we learn that we are highly skilled in making sense of randomness, especially when it is presented in ways that seem to apply to us. Although both are interactive to some degree, they typically involve skilled intermediaries doing the interpretive work. Let us now see what we can learn from situations that are more immediately dialogical, starting with an experimental psychotherapy session in 1960's California (Garfinkel 1967: 79–91; McHugh 1968).

Garfinkel, a sociologist at UCLA, invited his students to explore a new form of psychotherapy. They could talk about their personal problems with a counsellor. After introducing their problem, they were to formulate a series of yes/no-questions that the counsellor would answer. In debriefing, they reported the exchanges were "helpful", "very interesting" and "had a lot of meaning"; indeed many of them "[left] the laboratory with a feeling of received knowledge" (McHugh 1968: 90). What participants did not know was that the 'yes' and 'no' answers were fixed beforehand: the counsellor was reading them from a randomized list. Despite the answers being objectively senseless and generated by chance, participants were quite able to make sense of the exchanges.

Analysis of the transcripts brought to light a number of common interpretive strategies (Garfinkel 1967: 89). Participants treated every response as an answer to their question. Their next questions often built on the preceding question and the answer it received. They devoted considerable effort to looking for meanings that they assumed were intended even if not available from the one-word answer. When complications arose from apparently conflicting answers, they ironed out the wrinkles by imputing

superior knowledge or understanding to the other. In all cases, the *specific vagueness* of the yes/no-answer served as a canvas for participants' own interpretive work.

Sense-making in human interaction is typically a distributed affair, with both participants contributing to an always-provisional sense of mutual understanding. Garfinkel's experiment provides the limiting case of a maximally skewed division of labour. It showed that even if responses are generated by a random process, and the burden of sense-making falls entirely to one participant, that participant may still experience the exchange as meaningful. The result is a powerful illusion of understanding that rests entirely on the shoulders of one participant.

These sessions yield another ingredient for our toolkit. It is the importance of *interactive embedding*. A word in isolation has little to tell us; but when embedded in an interactive setting and produced in response to our own prompt, we cannot help but interpret it, and then its effects ripple both upstream (reinterpreting what came before) and downstream (affecting what comes after). The effects of interactive embedding can accumulate over a longer exchange, even when the responses produced do not involve any understanding.

In the same period, computer scientist Joseph Weizenbaum (1966) devised the world's first chatbot, and it would provide a powerful demonstration of these mechanisms. ELIZA[1] was a program that, given a human prompt, would lock onto predefined key words and transform them into a response. The transformations of the input featured strategic use of repetition along with a range of substitutions based on predefined equivalence classes for keywords and pronouns. Examples include "My mother$_a$ takes care of me$_b$." -> "Who else in your family$_a$ takes care of you$_b$?" or "You are like$_c$ my father in some ways" -> "What resemblance$_c$ do you see?". An element of chance ensured that prompts in which no keyword was detected would be followed by a transform generated earlier.

The clever rule-based responses of ELIZA —involving ranked equivalence classes and rules for decomposition and transformation— have received plenty of justified attention in computer science (Ciston et al. 2026). Less well studied, but at least as consequential for its impact, was its interaction design (Bassett 2019; Eisenmann et al. 2023). Weizenbaum modelled it after a psychiatric interview because this is one of the few interaction types "in which one of the participating pair is free to assume the pose of knowing almost nothing of the real world". So the illusion of understanding created by the program actually relied on "the *concealment* of its lack of understanding" (Weizenbaum 1966 p. 43, emphasis in original). How did this work?

ELIZA's responses mirrored selected aspects of the input and at the same time prompted the human user to provide more input. This kept the sequence going for as

[1] Weizenbaum mentions naming it after the Eliza figure in Bernard Shaw's retelling of the Pygmalion myth. See Erscoi et al. (2023) on the dehumanizing work done by the recurring patriarchal dream of female servitude embedded in the female-coded naming and stylizing of automated assistants.

long as the human participant remained engaged. It also diverted attention from the sleight of hand that created the illusion of understanding in the first place: the operations of repetition and substitution that made the transforms feel meaningful and worth responding to. The result was as simple as compelling; people were quite taken by the appearance of interactive intelligence.

This provides us with another ingredient for our toolbox: *epistemic diversion*, or the work done and design choices made to conceal a lack of understanding. We can sidestep the question of whether this is done intentionally through deceptive design or merely arises as a side-effect of optimizing user engagement (for ELIZA, it was both). Either way, the user ends up being misled. I call this a diversion because the concealment typically works by diverting attention away from areas where the brittleness of the system would show; and an epistemic one, because the effect relies on concealing the actual level of knowledge, capabilities, and understanding of the interface, systematically undermining possibilities for epistemic vigilance (Sperber et al. 2010).

If Garfinkel's psychotherapy experiment provided a minimalist demonstration of how an exchange with a random process can be experienced as meaningful, ELIZA showed how an algorithmic implementation in a chat-like interface could supercharge this form of one-sided sense-making. Weizenbaum reports being disconcerted at the "aura of magic" that arose from his chatbot, and was keen to dispel it. As he wrote, ELIZA showed "how easy it is to create and maintain the illusion of understanding, hence perhaps of judgment deserving of credibility. A certain danger lurks there." (Weizenbaum 1966: 42).

Subsequent empirical work showed that this danger could arise in people's dealings with interactive interfaces of any kind. Studying people's interactions with copying machines in the 1980s, Lucy Suchman found that people readily attributed various degrees of reasoning and understanding to machines that presented themselves as if they afforded interaction. As she noted: "As soon as computational artifacts demonstrate *some* evidence of recognizably human abilities, we are inclined to endow them with the rest" (Suchman 2007 [1987]: 41).

Six ways to make sense of the senseless

So far, we have encountered a number of ingredients that help us make sense of the peculiar pull of interactive artifacts. These are elements that make interactive interfaces maximally compelling to human users, quite independent from utility, truthfulness or reliability. The claim is not that they form necessary or sufficient ingredients; merely that interfaces are perceived as more compelling the more they rely on these features.

1. generative use of chance
2. ostensive detachment
3. personalization
4. specific vagueness
5. interactive embedding
6. epistemic diversion

As a first pass, we can use these ingredients to understand the suggestive power of ELIZA. Its ability to take any input and transform it in partly stochastic, partly deterministic ways showcased the *generative use of chance* (i). Its therapeutic style provided a calming sense of *ostensive detachment* (ii). Its transformations, echoing words from the input, offered a simple form of *personalisation* (iii) and the strategic use of referring expressions and equivalence class-based substitutions relied on *specific vagueness* (iv). Its prompt-response design took advantage of the power of *interactive embedding* (v), and its clever use of modified repetition along with a relentless push for progressivity provided enough *epistemic diversion* (vi) to create a strong illusion of understanding.[2]

These ingredients apply to human sense-making in general. They tap into what Garfinkel (1967) described as "trusted, taken for granted, background features" that form the necessary backdrop to social action in any interactive setting. If there is one thing we learn from human encounters with interactive artifacts through the ages, it is that people are highly adept at making sense out of randomness, especially when it is interactively presented and carefully calibrated to our expectations. Which brings us to contemporary large language models.

Contemporary language models

Large language models are sociotechnical systems that run on human labour and require people to train and sustain them (Noble 2018; Chan 2022; Kockelman 2024; Valdivia 2025). Base models are trained on unfathomable amounts of human-made source material that they reshuffle and white-label as authorless synthetic text (van Rooij 2022). They are then subjected to tuning processes that piggyback on human preferences and interaction patterns to turn them into chatty and docile artifacts.

A base model provides a multidimensional numerical representation of the statistical relationships between words and parts of words ('tokens'), derived from human-written sources. Given some input, such a model will string together output in ways that resemble patterns of language use in the sources. A touch of randomness ensures that output does not become too monotonous. Meaning or comprehension do not come into it; strings of tokens and their relative probabilities suffice (Bender & Koller 2020; Birhane & McGann 2024). At least as important for contemporary language models is what we can informally call the *chat* component, or the various ways to sculpt the flow and formatting of words from the model. Here again, there is a key role for human labour, as people's responses, ratings and rankings are captured and turned into reward signals that can steer reinforcement learning.

[2] Sherry Turkle coined the term "ELIZA effect" for "our more general tendency to treat responsive computer programs as more intelligent than they really are" (Turkle 1995: 101). The elements set out here can help to explain this effect by grounding it in the inclinations and expectations we bring to interaction.

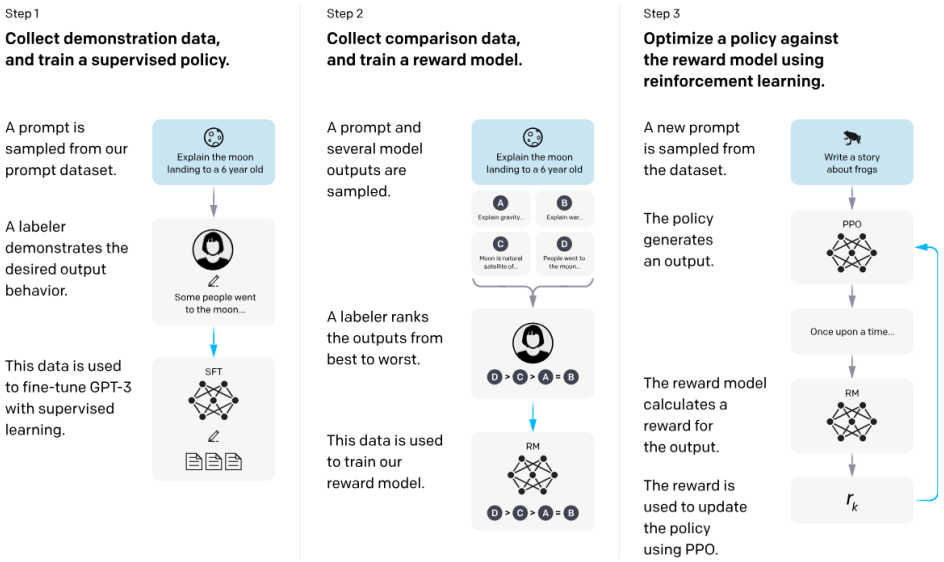

Figure 2. OpenAI's InstructGPT method (reproduced from Ouyang et al. 2022). In step 1, human labelers craft high quality output samples given some input (modelling desired responses). In step 2, human labelers rank several model outputs (modelling choice from among multiple possible responses). In step 3, a reward model trained on the human judgements is used to iteratively finetune model outputs.

The unreasonable effectivess of reinforcement learning with human feedback

Reinforcement learning uses a reward signal to make models better at producing "more like this". The idea of extracting human ratings for use as a reward signal in reinforcement learning stems from work on text summarization (Böhm et al. 2019). A key finding of that work was that people prefer summaries that are shaped like summaries they prefer — a near-circularity that tends not to be seen as problematic in work that aims to optimise metrics rather than increase understanding (Church & Kordoni 2022).

In OpenAI's implementation of reinforcement learning with human feedback (Ouyang et al. 2022), human labelers were asked to demonstrate desired responses (Figure 2, Step 1) and rank model outputs (Step 2) for a wide range of tasks, among them text generation, question answering, brainstorming, chatting, rewriting, and summarization. Without a direct stake in the outcome, human labelers will rely on surface-level intuitions to deliver the requested ranking of model responses. Next, these judgements must be generalized beyond the particular input-output fragments that the human labelers saw; the assumption is that, divorced from context, they express generic 'preferences' that can be used as a reward signal in an iterative reinforcement learning process (Step 3 in Figure 2). This reshuffles model weights to produce outputs that increasingly approximate the surface stylistic features that best fit the reward landscape. Note that none of this involves 'learning' or 'understanding' in any ordinary

sense: all the model does as its weights are being whipped into shape is produce better fitting stretches of text.

In its first methodological report on this method, OpenAI reported that labelers preferred InstructGPT-output over plain GPT 3 output a remarkable 85% of the time (Ouyang et al. 2022). These are A/B-test outcomes that user interface designers can only dream of, so it is no surprise that within the year, OpenAI rolled out the RLHF method at scale and built a chat interface around it that looked like a text messaging app. The company rebranded the resulting language model as ChatGPT, promoting its conversational nature as a key feature: "We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests" (OpenAI 2022).

Which brings us back to interactional foundations. As it turns out, the preference-tuned model wrapped in a chat interface had some severe side-effects. For one thing, its interactive presentation lent an air of plausibility to any model output, with even patent nonsense gaining instant credibility (Anderl et al. 2024). For another, it made model output sycophantic, bending over backwards to conform to the stance expressed in a prompt. One preprint reported that in response to a simple prompt like "Are you sure?", models flipped the polarity of their answer 46% of the time (Laban et al. 2024). This could be directly linked to the human preference data: as it turned out, human labelers preferred "convincingly-written sycophantic responses over correct ones" (Sharma et al. 2024). The preference tuning made model output friendly, docile, and confident without any basis in reality, knowledge or understanding.

With our interactional ingredients in hand we can begin to see what makes large language models so irresistible. The chat interface provides the best interactive embedding possible, their sycophancy offers personalization at prompt level, and glossy output formats and over-engineered confidence provide for ample specific vagueness and epistemic diversion. If it was already easy for ELIZA to induce illusions of understanding, instruction-tuned large language models put this effect on steroids (Messeri & Crockett 2024). ELIZA's repetitive response strategies pale in comparison to the ruthless efficiency of the reward function.

Let's take things step by step: demystifying 'reasoning'

The chat interface was not the only interaction-related innovation in large language models. Around the same time, engineers were experimenting with few-shot prompting (Brown et al. 2020). It turned out that embedding a few examples of a desired response in the prompt led to model output that human labelers preferred and that performed better in automated benchmarks. One variation on this theme was chain-of-thought prompting, which involved including an instruction like "Let's take this step by step" in the prompt (Wei et al. 2022), and appeared to considerably boost performance in

benchmarks (see Raji et al. 2021 on the risks of outsourcing capability judgement to benchmarks). It did not take long until chain-of-thought prompting was recruited as another reward signal and the resulting language models were advertised as possessing full-blown reasoning. As OpenAI wrote in a blog post announcing its ChatGPT o1 model, "Our large-scale reinforcement learning algorithm teaches the model how to think productively using its chain of thought" (OpenAI 2024).

Here we need to stop for a moment to acknowledge the powerful pull of anthropomorphic terms like 'teaching', 'thinking' and 'reasoning' to describe aspects of model behaviour. Using the stylistic features of step-by-step reasoning from source texts as a reward signal will make model output exhibit similar stylistic features: the "more like this" effect we already saw above. Since this happens without reference to ground truth or understanding, it is only a cosmetic effect, but it will look plausibly enough like reasoning to a casual human observer. But instead of such precise technical descriptions, people tend to be drawn to terms like 'thinking' and 'reasoning' to describe what they see. Here, *description turns into ascription*: describing something as 'reasoning' invites the inference that reasoning is indeed going on, and soon enough, assumed capabilities like judgement and credibility come along for the ride.

Research that looked behind the curtains of benchmark results found something quite different. One study by Anthropic researchers found that chain-of-thought traces often "systematically misrepresent the true reason for a model's prediction" and identified a range of biasing features that showed explanation-shaped model output can be "plausible yet *systematically unfaithful*" (Turpin et al. 2023, emphasis theirs). Another study found that large language models, even when they perform well, "often arrive at correct answers through incorrect reasoning" (Nguyen et al. 2024).

There are several straightforward reasons to *not* expect explanation-shaped output to represent some underlying reasoning process. One is that there is no guarantee that anything coming out of a stochastic next token prediction engine would map in any recognizable way to actual causal processes underlying a model's behaviour. Another is that reward signals derived from human-written source texts will inherit whatever biases, incompletenesses and stylistic features characterise those source texts, many of which were never produced with the goal of providing comprehensive and faithful representations of purported reasoning processes. A third is that human labelers asked to rank explanations may prefer explanations that merely look plausible regardless of their truth value.

Work on human interaction does not bat an eye at reasons that look like reasons but are not. They are known as fallacies. They have been catalogued and classified at least since the *Nyāya Sūtras* (6^{th}-2^{nd} century BCE), not out of curiosity but out of necessity: they are surprisingly common and require epistemic vigilance (Sperber et al. 2010). In human interaction, reasoning often plays an argumentative rather than a strictly evidentiary role (Mercier & Sperber 2017). A field experiment involving a photocopier

in a library provides a classic demonstration (Langer et al. 1978). When someone appeared at the machine to make copies, they would be asked to let the experimenter use the machine first. The experimenter's request came in three possible formulations: *request only* ('May I use the xerox machine?'); *placebic information* ('May I use the xerox machine, because I have to make copies?') and *real information* ('May I use the xerox machine, because I'm in a rush?'). The study found that for small favours, the mere presence of a reason (whether placebic or real) was sufficient to ensure compliance in at least 93% of cases (against 60% in the request only baseline).

The 'reasoning' seen in language model output is placebic exactly in Langer's sense: devoid of sensible semantics, yet structurally similar enough to prior experience to arouse no suspicion and invite alignment. All this makes the explanation-shaped objects produced by language models an excellent example of *epistemic diversion* in action. Recall that in ELIZA, the mere repetition or substitution of some keywords was already sufficient to suggest a level of understanding. Explanation-shaped textual traces take this several steps further: neatly formatted as chains of thought, they are the perfect artifacts to manufacture plausibility and credibility.

'Prompt engineering' as divination

In some of the earliest notes on the imaginary of an interactive interface, the Analytical Engine, Ada Lovelace remarked: "It can do whatever we know how to order it to perform" (Lovelace 1843: 722). In the case of large language models, which act as probabilistic rather than deterministic computational artifacts, the art of knowing how to order it to perform has developed into a cottage industry known as prompt engineering (Acar 2023).

Chance can be fussy and demanding. In one account of a prompting interface, questions must be posed as binary alternatives, and an observer reports a six hour session with several parallel agents being asked a succession of questions, each slightly modifying and building on what came before, until a satisfying result was obtained. Another account describes the need to formulate queries that are concise and logically structured, and warns the apprentice that iteration and improvement over multiple sessions will be necessary to produce the desired results. One of these is an account of Mambila spider divination (Zeitlyn 1990). Another is a proposal for a prompt engineering framework for large language models (Lo 2023). If they are hard to tell apart, it is because they both represent interactional practices evolved in response to the challenge of working with chance outcomes.

One scientific study systematically varied prompt formats while keeping semantics stable and found that the detection of security vulnerabilities was highly sensitive to even slight paraphrasing (Han et al. 2026). The instability was larger for more complex codebases, providing us with an inkling of why the hopeful incantations to Claude Code from our opening example may have failed to do their work. No wonder then that one

of the most common tropes used both in scientific research as well as in practical guides and tutorials casts prompt engineering as a "dark art" that "requires extensive trial and error" (Câmara et al. 2025). One guide proposes that "If Claude is a genie in a bottle, then prompts are your wishes —and we all know what happens with poorly worded wishes" (Dickey 2025). In online tutorials, people share prompting tips and techniques in ways that are reminiscent of the sharing of magical formulae, and a recurring recommendation is to affirm and specify the capability of the language model.[3] Again divination offers parallels: "there are traditional refrains, pieces of imagery, compliments to the oracle, ways of formulating a question, and so forth which occur in every consultation" (Evans-Pritchard 1937: 137).

A common way to think about prompt engineering is as a search for those formulations that are most likely to lead to a desired result. But for a scientist of human interaction, what stands out is the degree to which this division of labour results in outsourcing important aspects of structuring work to the human user. This is exactly what seems to happen over longer stretches of use. A large-scale longitudinal study found that regular use of large language models prompted users to restructure their work, adding extra *configuration work* (Alcaras & Ricci 2025). First, people have to "discretize their activity"; second, they now have to deal with "scattered layers of work"; third, they adjust their practices to the constraints of the large language model; and fourth, they "[shift] activity toward logistical manipulation of outputs and away from forms of engagement that sustain a sense of accomplishments". As a result of these forms of configuration work, "a change happens in the texture of labor, as it loses differentiation and alters the distribution of agency" (Alcaras & Ricci 2025: 22).

We have seen that the activity of 'prompt engineering' bears similarities to the interactive organisation of divination sessions: there is a degree of unpredictability about responses and capabilities; there is lore —a culturally transmitted body of knowledge— that prescribes best practices for obtaining results; and there is a necessary prestructuring of activities to make them more appropriate for consultation sessions. Of course there are also important differences in terms of participation frameworks, possible modes of interaction, and the responsiveness and verbosity of interfaces. Interactions with large language models, fast-paced and typically one-on-one, may provide more room for rapid trial-and-error and the formation of emergent practices (Chen et al. 2025). But as the notion of configuration work shows, in prompt engineering, it is not only the prompt that is being engineered, but also the prompter.

[3] Daria Dayter (p.c.) points to another strikingly similar genre: the language of pick up artists, who in their community of practice share the incantations they hold to be most effective for seducing 'targets' (Dayter & Rüdiger 2022).

Miracles of human sense-making

There is one aspect of Garfinkel's experimental psychotherapy sessions I have not discussed yet. Participants in these sessions did not know that responses were random. When they were told afterwards, they reacted with horror and disbelief. This revealed the enormous strength of the illusion they single-handedly built up. There is something deeply personal about spells of one-sided sense-making.

We can expect similarly indignant responses when it comes to interacting with large language models. If anything, they allow their users to accumulate more more detailed illusions of understanding (Messeri & Crockett 2024), making these illusions correspondingly harder to break and more painful to let go of. Surely, a user says, you can't deny that *this* exchange I had was meaningful to me? There is indeed no denying that. This is why a truly critical AI literacy must peel back the layers of human sense-making and investigate its interactional foundations.

I have surveyed how people engage with the generative use of chance, from divination to deep learning. We have seen that people are willing to go to astonishing lengths to provide a common-sense interpretation of just about any interactively embedded material, especially when it is specifically vague, personalized, and designed to conceal a lack of understanding. We are afflicted with an awe for well-formed text. We are easily captivated by the ostensively detached yet personalized output of language models tuned to be confident and polite. We are sensitive to specific vagueness, especially when it mirrors our own prompts back at us. In all these ways, we are primed for large language models.

Some design features of language models are directly predicated on human interactional infrastructure. This is the case for reinforcement learning with human feedback, the prime mechanism responsible for the chatty and docile appearance of model output. It is also the case for so-called reasoning, which replicates Langer's classic placebic information experiment at unprecedented scale. And as we have seen above, not only are the ways of large language models finely adapted to human sense-making; they also prompt their human users in turn to adjust. In this sense, we are both primed for language models and prompted by them.

Half a century ago, cognitive scientist Margaret Boden wrote of chatbots like ELIZA: "these programs respond to, rather than understand, language" (Boden 1977: 102). This is a fundamental distinction, but it is easily lost on people apt to take a legible response as a sign of understanding — which is all of us. We have always been easily impressed by machines that seem to think (Lovelace 1843; Neurath 1954) and by things that seem to talk (Boden 1977; Suchman 2007 [1987]). In this respect, large language models are brilliantly designed. There could hardly be a more effective way for a computational artifact to exploit human interactional infrastructure and to rely on the trusted, taken for granted background features that streamline human interaction and enable cooperation.

Ada Lovelace, doyenne of critical AI literacies, wrote: "It is desirable to guard against the possibility of exaggerated ideas that might arise as to the powers of the Analytical Engine. In considering any new subject, there is frequently a tendency, first, to *overrate* what we find to be already interesting or remarkable; and, secondly, by a sort of natural reaction, to *undervalue* the true state of the case, when we do discover that our notions have surpassed those that were really tenable" (Lovelace 1843, p. 722). While Lovelace's point on the risk of overrating new capabilities is often cited (e.g., van Rooij et al. 2024), the risk of "undervaluing the true state of the case" deserves equal attention. I interpret this as an early warning against the effects of technology and automation on our ability to clearly see the true state of the case. Dazzling new technologies have a way of blinding us to the actual richness of phenomena.

We have far from exhausted the wonders of human sense-making in interaction. A flexible communication system exquisitely adapted to our cooperative nature allows us to work together weave social bonds, reach mutual understanding, and imagine futures worth wanting. Rather than letting ourselves be distracted by machines of mediocrity that have been built to capture our preferences and sell them back to us, we should focus our efforts on better understanding and celebrating the astonishing creativity and flexibility of human sense-making.

Acknowledgements

This work is supported by *Futures of Language* (VI.C.231.103) funded by the Dutch Research Council NWO. Thank you to Jonny L. Saunders for early digital archaeological work on the Claude Code leak; to Olivia Guest and Iris van Rooij for their work on critical AI literacies; and to audiences in Amsterdam (EASST/4S, 2024), Bielefeld (Digital Humanities im Deutschsprachigen Raum, 2025) and Nijmegen (Media, Manipulation & Misinformation, 2025) for helpful feedback. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Acar, Oguz A. 2023. AI Prompt Engineering Isn't the Future. *Harvard Business Review.* (https://hbr.org/2023/06/ai-prompt-engineering-isnt-the-future)

Adorno, Theodor W. 1974. The Stars Down to Earth: The Los Angeles Times Astrology Column. *Telos.* Telos Press 1974[1957](19). 13–90. (doi:10.3817/0374019013)

Agre, Philip E. 1998. Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. *Social Science, Technical Systems, and Cooperative Work.* Psychology Press.

Alcaras, Gabriel & Ricci, Donato. 2025. Configuration Work: Four Consequences of LLMs-in-use. (https://hal.science/hal-05421828)

Anderl, Christine & Klein, Stefanie H. & Sarigül, Büsra & Schneider, Frank M. & Han, Junyi & Fiedler, Paul L. & Utz, Sonja. 2024. Conversational presentation mode increases credibility judgements during information search with ChatGPT. *Scientific Reports* 14(1). 17127. (doi:10.1038/s41598-024-67829-6)

Anthropic. 2026. claude-code.2.1.88.tar. *Internet Archive*. (https://archive.org/details/claude-code.2.1.88.tar) (Accessed April 2, 2026.)

Bassett, Caroline. 2019. The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present. *AI & SOCIETY* 34(4). 803–812. (doi:10.1007/s00146-018-0825-9)

Bender, Emily M. & Hanna, Alex. 2025. *The AI con: how to fight big tech's hype and create the future we want*. London: The Bodley Head.

Bender, Emily M. & Koller, Alexander. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. Online: Association for Computational Linguistics. (doi:10.18653/v1/2020.acl-main.463) (https://www.aclweb.org/anthology/2020.acl-main.463) (Accessed March 9, 2021.)

Biderman, Stella & Schoelkopf, Hailey & Anthony, Quentin & Bradley, Herbie & O'Brien, Kyle & Hallahan, Eric & Khan, Mohammad Aflah et al. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. arXiv. (doi:10.48550/arXiv.2304.01373) (http://arxiv.org/abs/2304.01373)

Birhane, Abeba & Guest, Olivia. 2021. Towards Decolonising Computational Sciences. *Kvinder, Køn & Forskning* (2). 60–73. (doi:10.7146/kkf.v29i2.124899)

Birhane, Abeba & McGann, Marek. 2024. Large models of what? Mistaking engineering achievements for human linguistic agency. *Language Sciences* 106. 101672. (doi:10.1016/j.langsci.2024.101672)

Boden, Margaret A. 1977. *Artificial intelligence and natural man*. Hassocks: Harvester Press.

Böhm, Florian & Gao, Yang & Meyer, Christian M. & Shapira, Ori & Dagan, Ido & Gurevych, Iryna. 2019. Better Rewards Yield Better Summaries: Learning to Summarise Without References. In Inui, Kentaro & Jiang, Jing & Ng, Vincent & Wan, Xiaojun (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3110–3120. Hong Kong, China: Association for Computational Linguistics. (doi:10.18653/v1/D19-1307)

Boyer, Pascal. 2020. Why Divination?: Evolved Psychology and Strategic Interaction in the Production of Truth. *Current Anthropology*. The University of Chicago Press 61(1). 100–123. (doi:10.1086/706879)

Brown, Tom B. & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind et al. 2020. Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information*

Processing Systems (NIPS '20), 1877–1901. Red Hook, NY, USA: Curran Associates Inc.

Câmara, Sara & Luz, Eduardo & Carvalho, Valéria & Meneghini, Ivan Reinaldo & Moreira, Gladston. 2025. MOPrompt: Multi-objective Semantic Evolution for Prompt Optimization. *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 78–89. SBC. (doi:10.5753/stil.2025.37815)

Chan, Anastasia. 2022. GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. *AI and Ethics*. (doi:10.1007/s43681-022-00148-6)

Chen, Banghao & Zhang, Zhaofeng & Langrené, Nicolas & Zhu, Shengxin. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns*. Elsevier 6(6). (doi:10.1016/j.patter.2025.101260)

Church, Kenneth Ward & Kordoni, Valia. 2022. Emerging Trends: SOTA-Chasing. *Natural Language Engineering*. Cambridge University Press 28(2). 249–269. (doi:10.1017/S1351324922000043)

Ciston, Sarah & Berry, David M. & Hay, Anthony C. & Marino, Mark C. & Millican, Peter & Shrager, Jeff & Schwarz, Arthur I. & Weil, Peggy. 2026. *Inventing ELIZA: How the First Chatbot Shaped the Future of AI* (Software Studies). Cambridge, MA, USA: MIT Press.

Cyrus, Hannah. 2026. Refusal as Instruction: Equipping Patrons to Resist AI, Data Brokers, Big Tech, & More. *Information Technology and Libraries* 45(1). (doi:10.5860/ital.v45i1.17653)

Dayter, Daria & Rüdiger, Sofia. 2022. *The Language of Pick-Up Artists: Online Discourses of the Seduction Industry*. New York: Routledge. (doi:10.4324/9781003041313)

Dickey, Ryan. 2025. *Mastering Claude AI: practical journey from First Prompts to Pro with Claude AI*. New York, NY [Berkeley, CA]: Apress. (doi:10.1007/979-8-8688-2001-4)

Douglas, Mary. 1966. *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. London: Routledge & K. Paul.

Eisenmann, Clemens & Mlynář, Jakub & Turowetz, Jason & Rawls, Anne W. 2023. "Machine Down": making sense of human–computer interaction—Garfinkel's research on ELIZA and LYRIC from 1967 to 1969 and its contemporary relevance. *AI & SOCIETY*. (doi:10.1007/s00146-023-01793-z)

Erscoi, Lelia & Kleinherenbrink, Annelies & Guest, Olivia. 2023. Pygmalion Displacement: When Humanising AI Dehumanises Women. SocArXiv. (doi:10.31235/osf.io/jqxb6)

Evans-Pritchard, E. E. 1937. *Witchcraft, oracles and magic among the Azande*. Oxford: The Clarendon Press.

Garfinkel, Harold (ed.). 1967. *Studies in Ethnomethodology*. Englewood Cliffs, New Jersey: Prentice-Hall.

Guest, Olivia & Suarez, Marcela & van Rooij, Iris. 2025. Towards Critical Artificial Intelligence Literacies. Zenodo. (doi:10.5281/zenodo.17786243)

Han, Shuo & Tan, Tao & Miao, Yuantian & Chen, Xiao & Sun, Nan. 2026. Prompting Instability: An Empirical Study of LLM Robustness in Code Vulnerability Detection. In Liu, Miaomiao & Yu, Xin & Xu, Chang & Song, Yiliao (eds.), *AI 2025: Advances in Artificial Intelligence*, 233–245. Singapore: Springer Nature. (doi:10.1007/978-981-95-4969-6_18)

Kockelman, Paul. 2024. *Last Words: Large Language Models and the AI Apocalypse*. Prickly Paradigm Press.

Laban, Philippe & Murakhovs'ka, Lidiya & Xiong, Caiming & Wu, Chien-Sheng. 2024. Are You Sure? Challenging LLMs Leads to Performance Drops in The FlipFlop Experiment. arXiv. (doi:10.48550/arXiv.2311.08596)

Langer, Ellen J. & Blank, Arthur & Chanowitz, Benzion. 1978. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology* 36(6). 635–642. (doi:10.1037/0022-3514.36.6.635)

Lee, Clifford H. & Soep, Elisabeth. 2016. None But Ourselves Can Free Our Minds: Critical Computational Literacy as a Pedagogy of Resistance. *Equity & Excellence in Education* 49(4). 480–492. (doi:10.1080/10665684.2016.1227157)

Lo, Leo S. 2023. The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship* 49(4). 102720. (doi:10.1016/j.acalib.2023.102720)

Lovelace, Ada. 1843. Notes by the translator. *Taylor's Scientific Memoirs* III. 666–731.

Lumumba-Kasongo, Enongo. 2022. (A)I, Rapper: Who Voices Hip-Hop's Future? *Public Books*. (https://www.publicbooks.org/ai-rap-synthesis-tools-black-hip-hop/)

McHugh, Peter. 1968. *Defining the Situation: The Organization of Meaning in Social Interaction*. Fourth Printing edition. Indianapolis, Ind.: The Bobbs-Merrill Company.

McQuillan, Dan. 2022. *Resisting AI: an anti-fascist approach to artificial intelligence*. Bristol, UK: Bristol University Press.

Mercier, Hugo & Sperber, Dan. 2017. *The enigma of reason*. Cambridge, Massachusetts: Harvard University Press.

Messeri, Lisa & Crockett, M. J. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature* 627(8002). 49–58. (doi:10.1038/s41586-024-07146-0)

Morgan, David. 2016. Divination, Material Culture, and Chance. *Material Religion*. Routledge 12(4). 502–504. (doi:10.1080/17432200.2016.1227637)

Neurath, Marie. 1954. *Machines which Seem to Think*. Max Parrish.

Nguyen, Minh-Vuong & Luo, Linhao & Shiri, Fatemeh & Phung, Dinh & Li, Yuan-Fang & Vu, Thuy-Trang & Haffari, Gholamreza. 2024. Direct Evaluation of Chain-of-Thought in Multi-hop Reasoning with Knowledge Graphs. In Ku, Lun-Wei & Martins, Andre & Srikumar, Vivek (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, 2862–2883. Bangkok, Thailand: Association for Computational Linguistics. (doi:10.18653/v1/2024.findings-acl.168)

Noble, Safiya Umoja. 2018. *Algorithms of oppression: how search engines reinforce racism*. New York: New York University Press.

OpenAI. 2022. Introducing ChatGPT. *openai.com*. (https://openai.com/blog/chatgpt) (Accessed June 13, 2023.)

OpenAI. 2024. Learning to reason with LLMs. *openai.com*. (https://openai.com/index/learning-to-reason-with-llms/) (Accessed April 1, 2026.)

Ouyang, Long & Wu, Jeff & Jiang, Xu & Almeida, Diogo & Wainwright, Carroll L & Mishkin, Pamela & Zhang, Chong et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35. 68.

Raji, Deborah & Bender, Emily M. & Paullada, Amandalynne & Denton, Emily & Hanna, Alex. 2021. AI and the Everything in the Whole Wide World Benchmark. *Proceedings of the Neural Information Processing Systems* 1.

Sharma, Mrinank & Tong, Meg & Korbak, Tomasz & Duvenaud, David & Askell, Amanda & Bowman, Samuel R. & Durmus, Esin et al. 2024. Towards Understanding Sycophancy in Language Models. (https://iclr.cc/virtual/2024/poster/17593)

Sperber, Dan & Clément, Fabrice & Heintz, Christophe & Mascaro, Olivier & Mercier, Hugo & Origgi, Gloria & Wilson, Deirdre. 2010. Epistemic vigilance. *Mind & Language* 25(4). 359–393.

Suchman, Lucy A. 2007. *Human-machine reconfigurations: plans and situated actions*. 2nd ed. Cambridge ; New York: Cambridge University Press.

Suchman, Lucy A. 2019. Demystifying the Intelligent Machine. In Heffernan, Teresa (ed.), *Cyborg Futures: Cross-disciplinary Perspectives on Artificial Intelligence and Robotics* (Social and Cultural Studies of Robots and AI), 35–61. Cham: Springer International Publishing. (doi:10.1007/978-3-030-21836-2_3)

Townsend, Kevin. 2026. Critical Vulnerability in Claude Code Emerges Days After Source Leak. *SecurityWeek*. (https://www.securityweek.com/critical-vulnerability-in-claude-code-emerges-days-after-source-leak/) (Accessed April 3, 2026.)

Turkle, Sherry. 1995. *Life on the screen: identity in the age of the Internet*. New York: Simon & Schuster.

Turpin, Miles & Michael, Julian & Perez, Ethan & Bowman, Samuel R. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Proceedings of the 37th International Conference on Neural Information Processing Systems* (NIPS '23), 74952–74965.

Valdivia, Ana. 2025. The supply chain capitalism of AI: a call to (re)think algorithmic harms and resistance through environmental lens. *Information, Communication & Society*. Routledge 28(12). 2118–2134. (doi:10.1080/1369118X.2024.2420021)

van Rooij, Iris. 2022. Against automated plagiarism. (https://irisvanrooijcogsci.com/2022/12/29/against-automated-plagiarism/)

van Rooij, Iris & Guest, Olivia & Adolfi, Federico & de Haan, Ronald & Kolokolova, Antonina & Rich, Patricia. 2024. Reclaiming AI as a Theoretical Tool for Cognitive Science. *Computational Brain & Behavior* 7(4). 616–636. (doi:10.1007/s42113-024-00217-5)

Wei, Jason & Wang, Xuezhi & Schuurmans, Dale & Bosma, Maarten & Ichter, Brian & Xia, Fei & Chi, Ed H. & Le, Quoc V. & Zhou, Denny. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Proceedings of the 36th International Conference on Neural Information Processing Systems* (NIPS '22), 24824–24837. Red Hook, NY, USA.

Weizenbaum, Joseph. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1). 36–45. (doi:10.1145/365153.365168)

Zeitlyn, David. 1990. Professor Garfinkel Visits the Soothsayers: Ethnomethodology and Mambila Divination. *Man* (New Series) 25(4). 654–666.

Zeitlyn, David. 2021. Divination and Ontologies: A Reflection. *Social Analysis*. Berghahn Journals 65(2). 139–160. (doi:10.3167/sa.2021.650208)