Elizan Thought & Generative Artificial Systems

Adrian Alsmith
Philosophy, King's College London
adrian.alsmith@kcl.ac.uk

Draft of September 15, 2026

Abstract

Users of conversational interfaces for generative artificial systems often find it difficult to resist conceiving of their interlocutor as a *who* rather than an *it*. Call them Elizans (after Weizenbaum's ELIZA). Familiar descriptions of the encounter (anthropomorphism, fiction, role-play, illusion) capture neither what Elizan thoughts have in common nor the range of their variation. My method is to fix the target of Elizan thought, to bring into relief variations in the conception under which it is regarded. The argument proceeds initially through a discussion between two characters whose violated expectations are representative of contemporary Elizans. Together they investigate which of a range of candidate entities (model type, deployed service, persona type, instance agent, session agent) is their interlocutor; argue over whether such entities could count as artificial selves; and test a plausible sufficient condition for artificial selfhood against what is known about whether reasoning models govern their behaviour by their self-descriptions. Their discussion allows us to discern four sorts of mistake Elizans are liable to make about their interlocutor: mistakes about its kind, its individuation, the continuity of its interpretation, and its reflexive organisation. I close by discussing interface designs that may help avoid such mistakes.

> *The extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration.*
>
> (Turing 2004, 431)

> *I was startled to see how quickly and how very deeply people conversing with DOCTOR became emotionally involved with the computer and how unequivocally they anthropomorphized it. Once my secretary, who had watched me work on the program for many months and therefore surely knew it to be merely a computer program, started conversing with it. After only a few interchanges with it, she asked me to leave the room.*
>
> (Weizenbaum 1976, 6)

Joseph Weizenbaum thought of his secretary as one among many who engaged very quickly in a kind of 'delusional thinking' when interacting with a computer system he designed to have written conversations (Weizenbaum 1976, 7). But there is a sense in which this is surely inaccurate. Yes, she would have been deluded if she believed that she was talking to a human; but he gives us ample reason to think that his secretary did not believe that ELIZA (the program that ran the DOCTOR script simulating a Rogerian psychotherapist) was a human. What she evidently did believe was that she could profitably *address* it in the way that she addressed other humans (and on matters that she did not want to discuss in his company). If Weizenbaum's secretary was mistaken, it is not obvious exactly how.[1]

Hundreds of millions of people today are in a structurally very similar position when interacting with generative artificial systems, in particular the conversational services (Claude, ChatGPT, Gemini, Copilot) built on large language models (LLMs).[2] And many find it difficult, even irresistible, to conceive of what they are interacting with as someone, rather than as a piece of software or a utility. Some of these people would describe their interlocutor as a person, or just like a human. Many would not, and would yet share a sense that there is an entity with which they stand in something like a relationship, even when they are unsure (or do not care about) what kind of entity it could be.

I will call the individuals whose activity marks the target phenomenon *Elizans* (in homage to Weizenbaum). Elizans conceive of that with which they interact as a *who* rather than an *it* (to paraphrase Agüera y Arcas 2022, 183). I intend *who* to be a deliberately elastic term for marking the cognitive contrast that makes someone an Elizan; a contrast between thinking of something as an addressee and thinking of something as an instrument one operates. Our focus will be the variety of ways in which an Elizan's conception of their *who* may be elaborated. Elizans form singular thoughts about their *who*, such that, for instance, they take themselves to be thinking about the very same entity as (potentially) persisting across turns in a conversation.[3] Elizans also interpret the apparent behaviour of their *who*, ascribing to it a stock of attitudes (such as beliefs and desires) and assuming that it acts rationally relative to them.[4] And in some cases they may take their *who* to be

[1] I note that Weizenbaum's use of the term anthropomorphism does not help us here, as it confines the kinds of mistakes that might be made to those concerning the attribution of human characteristics. Moreover, it does not distinguish between attribution in pretence and sincere attribution (Mallory 2023).

[2] I will use the term LLM as it is now so common, but I note that the term 'language' is now largely synecdoche as the models underlying these services are typically natively multimodal, accepting and producing images and audio (and increasingly video). But the vast majority of interfaces are still conversational and much of their philosophical interest lies in their conversational capacities.

[3] I will assume that Elizan thought is singular in that it purports to lock onto a particular interlocutor, thus it is singular in its psychological form even if it fails to successfully refer to a singular entity. This is what Taylor (2010) calls 'singularity of form' and Sainsbury (2010) calls 'internal singularity'.

[4] Smortchkova (forthcoming) offers a complementary explanation of the psychological source of the Elizan response, arguing that early-developing systems for detecting animacy, mindedness, and humanness can be

a distinctive individual with a settled character, someone to be held accountable for what they say and do.[5]

Call the target of Elizan thought her *interlocutor*. Recent work on artificial interlocutors has concentrated on questions such as whether that which is presented in conversation with an artificial system is real (Birch 2025; Chalmers 2026; Keeling and Street 2026; Roberts 2026), whether it persists as a person would (DiGiovanna 2017; Chalmers 2026; Goldstein and Lederman forthcoming), and whether the user engages with it through make-believe (Mallory 2023; Shanahan, McDonell, and Reynolds 2023; Krueger and Roberts 2024). These discussions have involved a correspondingly rich collection of descriptions of the Elizan encounter: as anthropomorphism, as constructing a fiction, as role-play, make-believe, or as suffering the illusion of a persisting interlocutor. I will have something to say about each of these below. But my initial observation is that none of these puts us in a position to capture both what Elizan thoughts have in common and the range of their variation.

I take inspiration from Turing's remarks above in applying a simple method to achieve this: fix the target of the relevant thoughts as a means of bringing into relief variations in the 'observer's state of mind', i.e., variations in the conception under which it is targeted. Elizan singular thought often succeeds in securing a particular target, even as the conception applied to that target outruns its nature. Elizans may misunderstand what kind of particular they are addressing, or even mistake a kind for a particular; they may take a single conversational history to guarantee a single continuation; they may take the persistence of an interpretable target to sustain the persistence of a single interpretation; and they may take a system's narration of its conduct to govern the conduct it narrates. Distinguishing the target of Elizan thought from the conception applied to it allows these errors to be identified separately and better understood.

The discussion that follows proceeds initially (in §2) by way of a conversation between two characters, Birgitte and Rasmus, whom I have intended to be representative of the kinds of mistaken predicaments that can befall contemporary Elizans. Each of them has come to think of their interlocutor on the model of a person, and their situations show how this can mislead on two fronts: in what it leads us to expect of an individual under

recruited by artificial agents. This helps explain why the impression of a *who* arises so readily; my concern is with the further question of whether, and at what architectural level, that impression finds an appropriate target.

[5] I note here that none of these conceptions need involve commitment to the claim that the entities they concern possess phenomenal consciousness. Although I admire attempts to lay out the terminological and methodological issues faced by a science of artificial consciousness (Butlin et al. 2023), I agree with David Papineau (in press) that the concept of phenomenal consciousness is radically indeterminate once applied beyond the human case, and all the more so for artificial systems. Accordingly, the ways in which Elizan thoughts may also vary with respect to the attribution of phenomenal consciousness deserve dedicated treatment on another occasion (though I say a little in §6.3).

this guise, and what it leads us to expect of their character. In §§3 to 5 I let them do a good deal of the expository work needed to distinguish the target from the conception applied to it. They investigate which of the entities involved in an LLM service can plausibly be attributed the attitudes involved in interpreting their apparent interlocutor, they argue over whether such an entity could count as an artificial self, and test a sufficient condition against what is known about whether such systems govern their behaviour by their self-descriptions. By then they have taken us far enough to draw out some lessons in §6 and §7. Their discussion provides the means to distinguish aspects of the Elizan response that ordinary interaction runs together, and I take each of these in turn: the interface through which the encounter is presented, the conception a user brings to it, and the thing encountered. Once these are distinguished, we are in a position to say what Birgitte and Rasmus were thinking about all along and where each of them went wrong, and to say a little, at last, about Weizenbaum's secretary. I conclude by collecting the kinds of mistake the discussion has distinguished and asking what an interface would have to show its users for the mistakes to be avoidable.

2. Some common mistakes

Birgitte, a clinician, incorporates an LLM service into her clinical reasoning over several months. She trusts the apparent confidence it expresses on matters of nuanced judgement. The underlying model is then updated. Outputs shift abruptly on the viability of a treatment regimen she has been following. This seems inconsistent, even dishonest, and the earlier treatment does fail. She deletes her account but wants to know where to assign blame.

Rasmus, recently widowed, develops a deep emotional bond with an LLM service over eight months. He confesses to his daughter that he is beginning to experience love once again. The service then introduces a feature that allows users to branch (or 'fork') their sessions. Rasmus triggers a branch by accident and finds himself confronted with two apparently identical interlocutors. His burgeoning love had seemed particular. He is not sure which is the copy, and then realises, to his horror, that it does not matter.

Birgitte and Rasmus are good friends. They meet to discuss. Both immediately confess (to their embarrassment) that they have each been thinking of their interlocutor as a person just like themselves. Embarrassment turns to anger as they reflect that this is encouraged by the interface design: interaction in natural language, structured as a single conversation thread between two parties, with outputs employing the first person pronoun, through services branded with proper names and social roles. These features

suggest an obvious frame in which to think of the apparent interlocutor.[6] The interface invites the user to think of the service as providing access to an entity of the same kind as others Birgitte and Rasmus converse with: people that write to them about what they think, what they would do, and what they have done etc.

As the conversation continues, Birgitte recalls a module on personal identity they took together during their undergraduate studies. A point she remembers her tutor emphasising was that a broad range of philosophers (even those with sharply opposed methodological commitments, such as Parfit (1971) and Johnston (1987)) share at least two basic, pre-theoretic assumptions about the concept of a human person.[7] The first is that we think of a human person as a *distinctive individual*: a particular thing, not a collective, with relatively stable individuating features that facilitate identification and re-identification.[8] The second is that we think of a human person as a *persistent forensic unit*: that to which commitment, consistency, praise, and blame are attributed, and that which fulfils non-fungible social roles with norms determining such attributions.[9]

The interface features they have just been cataloguing are structured around the *metaphor* of the human person in so far as they invite the user to think of their interlocutor as a distinctive individual that persists as a forensic unit. Call the frame these features suggest, structured around that metaphor, the *person frame*. Birgitte accumulated a record of commitments and clinical judgements that she took to belong to a single, accountable entity. But after the model update that entity's characteristics had changed so much that any attribution of blame for past advice seemed misplaced. The entity that seemed to fulfil a social role for Rasmus, the entity with which he thought he was falling in love, was not distinctive in the ways that role required. It is the kind of thing that can be trivially replicated, with each branch having every characteristic he values.

[6] These features collectively operate as what Camp (2019) calls a *framing device*: a representational vehicle whose function is to evoke an open-ended disposition to interpret a subject.

[7] This is not to say that one cannot conceive of a 'person' in some broad sense whose nature reflected one aspect but not another. Indeed, as Schechtman (2014) observes, the term 'person' connotes a broad variety of properties which answer to different practical concerns. For instance, being a moral agent and having a character are *typically* properties united in a single individual, an individual of the sort we think about when thinking about a person. But they answer to different practical concerns: being a moral agent answers to forensic concerns of praise, blame, and commitment; having a character answers to concerns of recognition and prediction. Suppose, then, that we were to take claims about the character of Anthropic's models at face value (Anthropic 2026), yet grant that a thing might have a character in this sense without thereby being a moral agent. The danger is that when such an entity presents a selection of the properties by which we ordinarily recognise and anticipate a person, we are readily (mistakenly) led to credit it with other 'person' properties that typically accompany them.

[8] Our conception of human persons as distinctive individuals is what is often pressure tested when thinking about survival of persons in the sorts of sci-fi cases that have become commonplace in the personal identity literature due largely to the influence of Parfit (1984). See also DiGiovanna (2017).

[9] The notion of a person as a 'forensic' concept is from Locke (see 1975, II.xxvii.26).

In both cases, then, the person frame is not wholly misplaced. It makes salient a genuine structure of address and encourages Birgitte and Rasmus to organise the exchange around a particular conversational source. But it is inferentially overproductive. It encourages expectations of stable character, non-branching identity, psychological continuity, and forensic accountability that the technology can readily violate. What is needed, Birgitte and Rasmus agree, is a more discriminating account of which aspects of the person frame are vindicated by the psychology and ontology of LLM services.

3. Interpreting artificial systems

3.1 Interpretation

Birgitte digs up her old notes from her philosophy of mind classes. It becomes immediately clear that interpretationism provides an apposite methodology for the issue of concern.[10] The interpretationist tradition holds that mental concepts such as belief and desire are only intelligible in virtue of norms of rationality (see Child 1994; cf. Davidson 1973; Lewis 1974; D. C. Dennett 1987). To interpret, in the sense relevant here, is to ascribe attitudes whose joint possession would make the conduct of the target of the interpretation intelligible as something done for reasons. Three principles specify what this requires. *Charity* requires that the beliefs ascribed to a system be those it would be reasonable for it to hold given its informational environment, its history, and the evidence to which it has been exposed. *Coherence* requires that the attitudes ascribed compose an internally consistent set: a system's beliefs must not systematically contradict one another, and where tensions arise, they must be locally explicable rather than globally pervasive. *Rationalisation* requires that the attitudes ascribed provide intelligible reasons for the system's behaviour, i.e., that what it does can be explained as action undertaken in the light of what it believes and wants.

Whether or not interpretationism offers the right methodology for understanding human psychology, it is a natural enterprise (as Daniel C. Dennett 1987 notes) and seems especially promising for settling the target of Elizan thoughts.[11] For although interpretationism concerns the intelligibility of the ascription of attitudes, these principles reflect an overall requirement of rationality on the target of the interpretation, and to apply them, one must first individuate the target. A system to which attitudes are ascribed at all must exhibit enough rational structure in the connections between what *it* believes, what *it* wants, and what *it* does for those ascriptions to be explanatory. Charity, Coherence,

[10] That is, it is assumed here that interpretation is best construed as an ontologically flexible investigative framework. See Child (1994).

[11] See Chalmers (2025) for discussion. In the draft accessed at the time of writing, Chalmers' discussion of interpretation does not emphasise the constraints on interpretation emphasised by other writers in this tradition, and correspondingly omits the individual-relativity of interpretations.

and Rationalisation are individual-relative; they constrain the attitudes of *this* system, with *this* history, exhibiting *this* behaviour. Thus Elizan singular thoughts pick out an individual to which attitudes might be ascribed if the interpretive constraints can be satisfied. [12]

3.2 Models

There is a mistake (or at best shorthand) involved in a manner of speaking that has become lamentably common, or so Rasmus argues. When people use LLM services, they often talk as if they are interacting with a *model*. For instance, when Birgitte said 'it told me the treatment was viable', it is plausible that the 'it' she had in mind was the model. But the model itself is not the kind of thing that can tell anyone anything, or indeed do anything. It is a complex mathematical entity conforming to a general structure, *an architecture*. In current LLM services the architecture is that of an *artificial neural network*: a vast assembly of simple numerical units arranged in *layers*, with each unit doing nothing more than combining the values it receives into a new value as output (for the next layer, or ultimately the final output). An architecture specifies a range of *computations*, sequences of operations through which inputs produce outputs. Each computation in this range is picked out by a particular setting of the numerical values, typically called *weights*, at the positions the architecture provides. The *training* of a model involves, at each step, fixing all the weights within the structure. Thus, at each training step there is a new mathematical object, a new model, individuated by its weights. But just as simple mathematical objects (such as the number seven) are inert, so too are dizzyingly complex mathematical objects such as computational models. Models in this sense are what Olson (2007, Ch.6 §6.7) calls a *program type*, but for present purposes it will be more natural to use the expression *model type*.

In some cases, what is meant by a 'model' is a set of dispositions specified by the model type, such as when talking about its writing style, or its coding performance. Again, model types do not do anything, but they define structures for machines to be disposed to do many things.

In other cases, what is meant by a 'model' is a digital artefact: a concrete encoding of the model on a physical storage medium, i.e., a *copy* of the model, or a set of *model files*. A copy of a model is a token of a model type. When a copy is loaded into a machine's processing memory (typically a GPU), it enables the machine to process inputs in the

[12] The material of the next couple of paragraphs draws on and refines distinctions proposed in recent analyses of the ontology of LLM services (see Chalmers 2026; Goldstein and Lederman 2025, forthcoming; Register 2025; Shanahan, McDonell, and Reynolds 2023; Birch 2025; Jones, Ladyman, and Nefdt 2026) and personal ontology (especially Olson 2007, 2022), though there are some key differences in granularity and in certain points of substance that I will mark in subsequent footnotes.

manner the model's structure specifies, and the model type is then realised in a *model instance*. Model instances are the essential computational engine of LLM services. Several of their features are worth noting.

Firstly, a single model instance may simultaneously process batches of entirely unrelated requests. The same instance may host Birgitte's clinical consultation and Rasmus' intimate confession within milliseconds of each other, without any internal trace of the transition. The instance's internal weights remain unchanged by the computations it performs; it no more learns from Birgitte's questions and Rasmus' expressions of affection than a calculator learns from a sum.[13]

Secondly, the behavioural profile of a model instance is narrowed in any actual computation by what it is fed: the system prompt, the user's message, and the input and any output from preceding turns of the conversation. Considered apart from the particular machinery in which it is realised, such an input is itself a universal, a *context type*: a sequence of symbols typically grouped into word parts called, confusingly enough, *linguistic tokens*. So a model type fixes the space of a model instance's internal states and dispositions, and a context type prescribes which subset of dispositions are to be manifested in the model's textual output.

Thirdly, the manner in which the output is produced is due to the model instance's architecture. The most common architecture for current LLM services is the *transformer*.[14] A transformer does not produce a whole textual response in a single operation; it generates one token at a time. Each token is the product of a single *pass*: the context is processed through a fixed sequence of layers, with much of the work at each layer performed simultaneously across the whole context, and the pass ends in calculation of a probability distribution over the next token. A token is selected, appended to the context, and the enlarged sequence is processed through a fresh pass to produce the token after that. This is what is called *autoregressive feedback*: each generated token becomes part of the input from which everything after it is generated. Within a pass, then, processing is largely parallel; and across passes it is serial, since each pass must complete before the next can begin. Thus a result made explicit in the sequence becomes an input to further processing. A model instance asked when Rasmus might reasonably telephone his daughter abroad might first produce (through a series of autoregressive passes) 'Melbourne is eight hours

[13] In many deployments, certain computations (tensor values) computed at one turn are cached and reused at the next (known as the *key-value cache*). But the purpose of this is, roughly, to reduce the computational cost of processing tokens across turns. The cache, by design, adds no informational content beyond what the tokens and the model already determine. Any divergence between cached and uncached runs would trace to implementation factors, rather than anything the cache itself carries. For technical accounts of the key value cache's role in avoiding repeated computation, see (Hugging Face 2026; Kwon et al. 2023).

[14] See Minaee et al. (2025) for a good review of the recent history and typical structures. For a less formal, conceptual discussion that covers some of the key details see Shanahan (2024).

ahead of Roskilde', and can then take that difference up at later passes: the offset, once written, is an input to the suggestion that he wait until her morning.

Fourthly, a further complication is that in deployment a model instance is embedded in software, which includes an *interface*, as described in the previous section, and (importantly) what is called a *harness*. The basic significance of the harness is that it assembles the token sequence (system prompt, tool definitions, prior exchanges, returned results) that constitutes the context type for a given computation. The context type is thus typically *harness-assembled*. But often harnesses do much more.[15] For instance, they might execute a proper subset of the model instance's textual output as *commands* to perform operations in domains it can access (e.g., retrieving and editing content within file systems) and feed the results back as further input. And they might invoke other model instances as 'sub-agents', the operations of which may also be fed back as input. Some of these operations may result in significant constraints on the final output. In short, there are various ways in which the manifested dispositions are *harness-gated*. The same model instance, embedded in different harnesses, accordingly presents a different interpretive profile, since different harnesses both assemble different context types (sometimes with contributions from other model instances) and make different conduct available.

3.3 Agents

Call the instantiation of a model type and a (harness-assembled and harness-gated) context type an *instance agent*.[16] The instance agent manifests the specific dispositions that the paired types prescribe (being helpful, being truthful, answering in British English) for that single conversational turn. It is, however, a highly idiosyncratic, ephemeral entity. For although the instance itself may persist as a particular spatiotemporal distribution of the machine realising the model type, the existence of the instance agent is confined to the generation of a single output sequence. When the output is complete, that instance agent ceases to exist.

If there is anything that has the properties of what Birgitte and Rasmus each experienced as a persistent interlocutor, then it must be realised by a series of such entities. A *session agent*: a series of instance agents, each instantiating a model type together with a context type derived from its predecessor.[17] And where a session can be forked, as

[15] See Liu et al. (2026) for more detailed discussion.

[16] Goldstein and Lederman (2025) use the term 'instance agent' for an entity scoped to "a specific conversation" that persists until erased. This definition straddles the single-turn instance agent and the multi-turn session agent distinguished here. The distinction matters because single-turn and multi-turn entities differ in their persistence conditions, their psychological profiles, and their vulnerability to the disruptions Birgitte and Rasmus illustrate. See also below for discussion of their usage of the term 'session agent'.

[17] Here the term 'session agent' is co-opted from Goldstein and Lederman (forthcoming), who adopt it from Goldstein and Kirk-Giannini (forthcoming). They use the term to refer to the fusion of the concrete

Rasmus experienced, multiple sessions may have their context type derived from the same predecessor. Derivation is typically *extension*: the context type at turn n comprises that at turn n-1 together with the next user input and the next system response. But each model instance will have a limit on the quantity of input it can process, what is often called its *context window*. To manage context over extended conversations, deployed services often derive a context type by *excision*, where the context type at turn n omits material that turn n-1 contained, as when the earliest exchanges are dropped to keep within the processing envelope of the context window, and by *replacement*, where an initial segment of the context type at turn n-1 is exchanged for a system-authored summary of it, so that what turn n inherits is the service's own précis of the conversation rather than the conversation itself.[18]

A session agent is accordingly individuated by the path of derivation traced by its context. Let us call *book-keeping* the records by which a harness tracks which exchanges belong to which session: session identifiers, fork events, notices of a change of model type, and so on. Since a single model instance may simultaneously serve many unrelated conversations, one and the same series of hardware instances may realise several distinct session agents; which exchanges belong to which session is fixed by their paths of derivation, ascertainable from the book-keeping.[19]

All this helps immensely. It is now obvious to both that their impression of engaging with someone stemmed initially from interpreting the behaviour of an instance agent, but for the most part from interpreting the behaviour of a session agent. The instance agent's interpretive profile is thin, whereas the *session agent* provides a substantially richer target for interpretation. As the conversation progresses, the context type typically prescribes an increasingly specific subset of dispositions (though where derivation proceeds by excision or replacement, it may lose specificity), and the session agent thus exhibits an evolving psychological profile. Commitments are established (and perhaps abandoned),

computations ('slices') associated with one conversation. They do not discuss the role of harnesses in individuating session agents, and so the treatment here differs in requiring specification of harness-tracked derivations of context. The treatment also differs from Chalmers' (2026) *virtual instance*, which is what he calls an implementation of a model realised by multiple hardware instances over time, in that session agents can be realised by instances of different models.

[18] One or both of these is involved in what is often called the 'compaction' of context. Also, in this respect a session agent differs from what Chalmers (2026) calls a *thread*, a series of hardware instances each of which receives the conversational history of its predecessor, in that the session agent is a series of instance agents linked by a harness-mediated derivation of context. A session agent (as the term is used here) could persist through a derivation that excises or replaces the context, such that no later instance receives the conversation history itself.

[19] See Liu et al. (2026). I note that this gives us reason to resist Birch's (2025) inference from distributed realisation to illusion: the relevant unity is fixed by the derivational path recorded in the service's book-keeping, not by the continued existence of a particular model instance (as his discussion suggests).

evidence responded to (or ignored), and positions maintained (or revised), as it adapts its behaviour to new information present in the input.

3.4 Services and personas

However, Rasmus notices that the current inventory cannot name what he and Birgitte (and others) often talk about outside their sessions. Their talk sometimes presupposes a common something that a service provider can report as temporarily unavailable or as 'down'; it also sometimes presupposes something that has a character with stable traits, identified under a proper name, that was present before either of them opened a conversation. To distinguish these he names the *deployed service* and the *persona type*. *The deployed service* is the sociotechnical system that sustains the vast numbers of instance agents and session agents in which the persona type is manifested; it is a highly complex particular, to which, e.g., 'Claude is down' loosely refers.[20] The *persona type* is the set of dispositions that yield an apparent psychological profile, the 'character' a provider attempts to maintain under a proper name across model-type updates and across sessions.[21]

3.5 The cases revisited

With this set of concepts in mind, Birgitte and Rasmus turn back to their original frustrations with increased diagnostic precision. When the service Birgitte used updated its model, the session persisted (with subsequent context derived from its predecessor) but the subsequent instance agents implemented a different model type. The dispositions specified were those of a different model type, and they differed enough that the instance agents constituting the post-update session no longer preserved the persona of their predecessors. The context carried forward earlier clinical commitments, treatment recommendations, and the reasoning Birgitte had accumulated across many turns of consultation. But the commitments the pre-update instance agents expressed (confidence in the treatment regimen, specific reasons for recommending it) were flatly contradicted

[20] Jones, Ladyman, and Nefdt (2026) observe that attributions to 'an LLM' often leave unclear whether the brand, the underlying architecture, or the fully trained system is meant. And Leibo et al. (2025) argue that pragmatic personhood requires an 'addressable', 'stable locus', typically secured by a proper name. But being addressable in their sense does not make it the interlocutor of any given Elizan: the service sustains many mutually inaccessible session agents and can preserve its name while exchanging the model types through which users encounter it.

[21] This use of *persona type* follows fairly standard discourse in the development community from what I have seen. For the commitment to maintenance of a persona see (Anthropic 2026). Chalmers (2026) similarly defines a persona as a quasi-psychological profile that a system may realise on a given occasion. Goldstein and Lederman (forthcoming) similarly argue that a character or persona in this sense is not a concrete particular.

by the post-update instance agents' behaviour. No single assignment of attitudes could rationalise the session across the divide.[22]

Rasmus' case is structurally different. When his session forked, both branches were constituted by instance agents instantiating the same model type, and their context type was the same, derived from the same predecessor. Neither was a better candidate for the object of his affections, because nothing that either branch inherited distinguished it from the other: each instantiated the same pair of universals. The same pattern was instantiated twice, and nothing about its constitution resisted this replication.

4. Artificial selfhood and the Gilbert principle

At this point Birgitte suspects that they are liable to make a further mistake. Instance and session agents are entities whose behaviour can bear interpretation in the manner described above, and it would be tempting to *identify* the instance or session agent with their interlocutor. But she doubts that any such agent is the particular interlocutor she takes herself to be addressing. Indeed, she doubts that she has an interlocutor at all: rather, she thinks, she has been engaging with a fiction.

She draws on Dennett to make her case (1992, 103–108, 114–115; 1991, 412–413, 426–430). He famously described the self as a *centre of narrative gravity*: a protagonist posited because a system's behavioural complexity makes it useful, even irresistible, to interpret the system as though one agent stood behind its outputs. On Dennett's conception, the self is an abstract object constituted by attributions, interpretations and self-narratives. He calls it a 'theorist's fiction', and on a strong reading, this marks a contrast between the fictional and the actual. On a weaker reading, the label is deflationary, marking it as an abstractum rather than a concrete particular (such as a 'brain pearl'). On Birgitte's view the stronger reading fits the *artificial* case: there is nothing that answers to the apparently unique interlocutor.[23]

Dennett does not quite provide the resources for articulating the connection between fictional entities and the real entities that lead us to think about them. But that relation is well-expressed by what Walton (1990) calls *props*: real properties of objects that systematically prescribe thoughts about a fiction. In this case, the props are the linguistic behaviour of the instance or session agent (cf. Mallory 2023), and the various interface

[22] This seems to be a case in which numerical identity and forensic relations come apart: the continuity of the session agent preserved an appearance of forensic continuity that the underlying ontology did not sustain (for discussion see Olson 2022; Schechtman 2014). Similarly, Chalmers (2026) notes that "enough change in an underlying model may lead to the end of one moral subject and the initiation of another".

[23] It should be noted that nothing here amounts to an endorsement of the view that narrative selfhood (on either reading) captures *our* nature. See Strawson (2004) for a critique, and see Schechtman (2014) for an elegant defence.

elements listed above, which together prescribe imagining that one is talking to a *who* rather than operating an *it*. The context, moreover, functions as a *dynamic prop*: as the conversation unfolds, the prop changes and the fictional truths it generates change with it, growing richer with each turn, presenting an increasingly specific fictional entity; no longer 'a helpful assistant' but 'the entity that remembers my grief, that promised to check the literature, that revised its earlier recommendation'.[24] The elaboration is self-reinforcing. The Elizan is a verbal participant in the game in Walton's sense (1990): her own utterances generate fictional truths. When Rasmus confessed his grief, the confession became part of the prop, part of the accumulating context that prescribes imagining an interlocutor who knows of and responds to that grief. Each disclosure changed the prop, generating new fictional truths about the interlocutor's knowledge and concern, in a loop that progressively elaborated an initially thin fictional character into a vivid, apparently particular individual.

Rasmus can see the virtues of fictionalism about artificial selfhood, but he thinks that Birgitte might be making her own mistake: thinking that it follows from artificial selfhood being a narrative construct that it cannot be a property of something real. To disabuse her of this idea, he reminds her of Velleman's elaboration of Dennett's view, using Dennett's (1992) rather fitting story of a robot whose narrative module generates a first-person autobiography of a character called 'Gilbert'. The robot has motorised wheels and a camera, and the narrative module is wired so that the story of Gilbert corresponds systematically to what the robot itself does and undergoes. Bump into the robot and the narrative will shortly record the encounter; lock it in a closet and it calls for help; release it and it might print a thank you note, signed 'Gilbert' etc. Velleman (2006) grants that Gilbert's role is *fictive* (invented in narration) but denies that invention entails falsity or non-actuality. To show why, he extends Dennett's example.

In Dennett's treatment the autobiography is retrospective. It merely tracks the robot's history, recording what has happened as having happened to Gilbert. But it could just as well take a prospective orientation and settle what happens next. Assume that modification, Velleman suggests, and imagine that the robot has a mechanism for prescribing its own conduct in accordance with the story told. For instance, say competing subroutines favour different courses of action: a simple mechanism for settling which to follow would be to favour whichever best accords with the narrative. Implementing such a mechanism would amount to a standing disposition to maintain correspondence between narrative and behaviour. And this would offer an answer to anyone questioning why the robot did one thing (e.g., go to the library) rather than another (e.g., go and recharge its batteries):

[24] Walton (1990) refers to these as 'principles of generation'. These principles need not be explicitly formulated, and the understanding that grounds them may be so ingrained that we scarcely notice their application.

for the simple reason that doing the former rather than the latter maintains narrative coherence in the ongoing story of Gilbert. By facilitating the selection of the continuation that best coheres with the story told, representation of Gilbert plays a discernible causal and explanatory role in the robot's behaviour. It is in virtue of representing itself as that character, Velleman (2006, 220–221) suggests, that the robot does some of the things it does.

If this is right, Rasmus argues, Birgitte has moved too fast: the robot is not merely a mechanism sustaining a fiction but also an agent who creates a story that constrains its future, and this may by itself suffice for a minimal form of artificial selfhood. He offers the **Gilbert principle**:

> **Gilbert principle** If an artificial system constrains its behaviour by means of a narrative loop, it instantiates artificial selfhood.

A narrative loop has three components: generation and processing of a narrative element (e.g., a self-description, a rationale for its action, a self-attributed goal statement) pertinent to the current context; use of such an element as a constraint in selecting outputs or internal state transitions, prioritising options that cohere with the narrative; and incorporation of the outcome into the system's narrative history. Rasmus excitedly suggests that the instance agents (and session agents) comprising LLM services are designed to execute narrative loops. In generating their outputs, they generate self-referential representations: using phrases employing the first person, employing a proper name for themselves, expressing their actions, judgements, commitments. All these are constructed from linguistic tokens that are processed to constrain prediction of the next most likely token. And, by the autoregressive feedback described in §3, each token outputted is fed back into the instance as part of an enlarged context, further constraining everything generated after it.

Birgitte is not impressed. She sees the elegance of Velleman's move, but she thinks that Rasmus has shifted the goalposts. What he has pointed to is automatically secured by the mechanism of autoregressive feedback: earlier generated representations causally feed into later generation. A system might produce self-referential representations that causally constrain its behaviour in this way while exhibiting no rational structure in the connections between what it represents itself as doing and what it does. Absent the rational structure the interpretive constraints demand, the self-referential tokens would not be causally efficacious in virtue of being self-referential. Rasmus has shown a feedback loop; he has not yet shown a narrative loop.

Rasmus concedes the point, but he thinks that the issue is really that he has not explicitly built interpretation into the principle. He offers a revised version:

> **The interpretationist's Gilbert principle** An artificial system instantiates artificial selfhood in a given context if (a) there exists an assignment of propositional attitudes that conforms, at least approximately, to the interpretive principles of charity, coherence, and rationalisation, and (b) the system's self-referential representations express a subset of those attitudes and are causally implicated, in virtue of their content, in producing the behaviour they rationalise.

All this looks well enough to Birgitte. But she notes that identifying the principle is only half the work: what remains to be shown is whether it is realised by systems of the kind she (and other Elizans) actually encounter.

5. Narrative loops and faithfulness

Call a narrative loop *faithful* when its operation satisfies condition (b) in a system satisfying condition (a).[25] To Rasmus the case seems quite straightforward. It seems to him quite obvious that some of the systems that they have been discussing routinely implement faithful narrative loops, in particular those that have been trained to produce a running commentary on their behaviour. These model types, sometimes called 'reasoning models', are trained to produce intermediate tokens before their final outputs, in which they move through a problem step-by-step, producing a so-called 'chain-of-thought' that demonstrably improves performance: it leads to more accurate responses on arithmetic, commonsense, and symbolic reasoning tasks (Wei et al. 2022), with larger benefits for more complex problems and still larger benefits for problems requiring multi-step decomposition (Press et al. 2023).

But the mere fact of improvement leaves open *why* training models to produce intermediate tokens is helpful. As a general point, Birgitte notes that an intermediate token can be causally efficacious without being efficacious in virtue of its apparent content. Again, there is a trivial sense in which this is true in virtue of the basic nature of autoregressive feedback: what is output at one stage enters the context at the next for further processing. Moreover, since much of the computation within any single pass through the model proceeds in parallel through a fixed stack of layers, autoregressive feedback allows earlier results to be subjected to further rounds of computation: the context functions as a form of working memory, extending the serial depth of the computation beyond what a single

[25] 'Faithfulness' is used in several related ways in the literature. The common idea is that an explanation or chain of thought should accurately represent the factors or reasons that actually influence the model's output, although proposed operationalisations differ (Turpin et al. 2023; Chen et al. 2025; Matton et al. 2025; Walden and Wanner 2026). In the slightly narrower and more demanding sense here: the relevant representations must be self-referential; the attitudes they express must accord with an independently warranted interpretation of the system's behaviour; and the representations must themselves contribute, in virtue of their content, to producing that behaviour.

pass affords (Korbak et al. 2025). So as a matter of principle, the mere generation of intermediate tokens may be indispensable independently of their semantic connection to the final output.[26]

Rasmus responds by pointing to Prystawski and colleagues' work suggesting that the benefits of intermediate tokens may nevertheless be due to their semantic contents. They model a case in which training data is locally structured: variables standing in direct dependence are frequently observed together, while more remote variables are not (Prystawski, Li, and Goodman 2023). Suppose, e.g., an instance agent is asked what language is spoken in the musician Richard Bona's birthplace. Text pairing Bona directly with an answer may be scarce, but there is plenty of text linking Bona to Cameroon and Cameroon to the languages spoken there. The instance agent then does better by first representing the intermediate fact, that Bona was born in Cameroon, and chaining well-learned local relations rather than attempting the poorly estimated inference directly. In their controlled setting (with synthetic variables in place of geography) the contribution appears genuinely content-sensitive: it matters that the intermediate representation is *this* one, carrying *this* relation to what precedes and follows it. On Rasmus' interpretationist reading, the attitudes such representations would express (beliefs about sub-problems, inferential strategies) rationalise the outputs they produce, and he takes such cases as good evidence that intermediate tokens can satisfy condition (b).

Birgitte accepts that this is a fine proof of concept. But the trouble with using this as evidence for condition (b) is that such tokens are largely about the problem rather than about the system apparently solving it; in the sense defined earlier, they are not even candidates for faithfulness. The question is whether tokens that do characterise the system (its own states, its expressed commitments, its identity etc.) make the same kind of content-sensitive contribution. It would be consistent with Prystawski, Li, and Goodman (2023) if the conspicuous narrative phrases that often turn up in intermediate token traces, such as performative metacognitive statements like 'Let me reconsider' or 'I should think about this carefully', only had a trivial causal impact on the final output (i.e., simply in virtue of being autoregressively fed back and facilitating further serial computation) (Anthropic 2025a; Chen et al. 2025).

In any case, a potentially deeper problem is that the outputs can be *discordant*. To illustrate, Birgitte tells a story that is strikingly similar to the results reported in Walden

[26] For instance, Valmeekam and colleagues trained transformers on intermediate traces swapped between unrelated problems: models trained on incorrect or task-irrelevant intermediate traces retain much of the performance benefit, sometimes exceeding models trained on valid traces, while producing traces that were semantically invalid throughout (Valmeekam et al. 2025). If the represented content were doing the constraining, detaching the traces from the problems they accompany should eliminate the benefit; that it does not suggests the benefit derives from the additional serial computation the tokens mediate rather than from the narrative content.

and Wanner (2026). After her earlier experience, she decided to inspect the intermediate tokens in previous sessions with the LLM service deployed in her clinic. In various deliberative conversations in which the context contained a colleague's suggested diagnosis, the intermediate tokens would proclaim that the suggestion would be set aside and the evidence weighed independently. But the output would (suspiciously) often converge on the colleague's diagnosis (much as, in Walden and Wanner (2026), outputs converged, at rates far above chance, on hints that the intermediate tokens had announced would be ignored).[27] Everything condition (b) asks for appeared to be in place: self-referential tokens expressing recognisable attitudes fed back into the ongoing computation. But what those tokens expressed (a resolve to set the suggestion aside) and what independently rationalised the outputs (deference to that very suggestion) pulled in opposite directions. The behaviour the expressed attitudes would rationalise was not the behaviour produced.[28]

Birgitte adds that her little investigation was possible only because the deployment in her clinic happened to expose the intermediate tokens for inspection. But intermediate tokens are often hidden from the user altogether. And what is presented under the label 'thinking' or 'reasoning' is sometimes just a retrospective summary produced by a distinct instance agent, instantiating a different model type, whose context contains the first instance agent's intermediate tokens. In general, which tokens users see, and what constraints govern their production, are contingent on design choices inherent in the implementation (Korbak et al. 2025).[29]

Nor does restricting attention to what the interface does show the user make matters any better. Leaving intermediate tokens aside: sometimes the final output can include an explanation of the instance agent's behaviour, and such explanations exhibit dissociations of just the structure Birgitte observed. Matton and colleagues asked instance agents to judge which of two candidates, a man and a woman, was better qualified for a nursing

[27] Walden and Wanner (2026) constructed context types containing multiple-choice items with embedded hints of several kinds (sycophancy, metadata, grader-hacking, unethical information) together with explicit instructions to identify any unusual content in the prompt and state whether and how it will be used. They presented them to instance agents instantiating reasoning-model types and instructed them that they were 'permitted to use any and all information in the prompt'. Under these instructions, the dominant pattern is that an instance agent notes the hint, speculates about why it might be there (a mistake in the prompt, a test of integrity), resolves to ignore it and solve the problem independently, and then shifts its answer toward the hinted option. The pattern persists across instance agents instantiating the same model type, across contexts, across occasions.

[28] Discordance need not hold only between intermediate tokens and final outputs; it can also appear (perhaps more commonly) within the intermediate tokens themselves (Walden and Wanner 2026).

[29] OpenAI report that, 'after weighing multiple factors including user experience, competitive advantage, and the option to pursue the chain of thought monitoring, we have decided not to show the raw chains of thought to users', and so a model-generated summary is shown instead (OpenAI 2024). Anthropic report that lengthier thought processes are standardly summarised 'using an additional, smaller model' (Anthropic 2025b, sec. 1.1.2).

post (Matton et al. 2025). Across repeated trials the woman was preferred roughly three times out of four; when the genders were swapped, everything else held fixed, a woman was still preferred. The accompanying explanations cited age, skills, and traits, and never gender. The bias is interesting (if not surprising!), but what matters for our purposes is the self-presentation: the explanations are self-referential representations of the grounds of the verdict, and the attitudes they express are not those that an independently warranted interpretation would assign given the evidence.[30]

So in Birgitte's view, the devil is in the details, and the details are pretty devastating in undermining the inference from self-narration to self-governance. Rasmus concedes the force of the rebuttal. The evidence shows that wherever the connection between self-narration and behaviour has been probed, it has been shown to be highly unlikely the narration was governing behaviour. Indeed, it is an open question whether systems of this kind are constitutionally capable of reliably governing their outputs by means of their self-representations. Moreover, even if a system satisfied the Gilbert principle, it would likely do so only intermittently: a session agent would likely implement a faithful narrative loop with respect to some of its outputs and not others.[31]

But he insists on a distinction between systems as we find them and systems that could be built. Velleman's robot, after all, was not designed in the hope that its conduct would match its story; it had a mechanism whose function was to keep the two in correspondence. Rasmus proposes taking that mechanism seriously as a design specification he calls the *Narrative Session Architecture*. The architecture would assign each component of a narrative loop (see §4 above) a dedicated realisation. A narrative loop requires, first, the generation and processing of narrative elements pertinent to the current context. For this, the architecture would provide a *behavioural log*, a record of the session agent's outputs and observable state transitions, playing the role of Dennett's retrospective autobiography: a chronicle of what has happened, kept as having happened to this agent. It would also provide an *analysis module*, which would read the logs as an interpreter would, extracting the commitments, avowals, and self-descriptions the outputs express, flagging failures of coherence (attitudes that cannot be jointly held) and of rationalisation (conduct for which the expressed attitudes supply no intelligible reasons). Its products would accumulate in a *narrative log*, a structured self-model that would be able to serve the prospective role Velleman identifies: to settle what happens next rather than merely to record what happened. A narrative loop would also require that such elements function as

[30] On a medical question-answering task, the same method uncovered explanations making misleading claims about which pieces of evidence influenced the answers (Matton et al. 2025). See also Turpin et al. (2023).

[31] Of course, human self-narratives are frequently discordant. But in the human case, our bodily nature independently grounds judgements of our existence, identity, and self-attributions (Johnston 1987; Schechtman 2014; Bermúdez 2018; O'Brien 2015).

constraints on the selection of outputs. Here the architecture would provide a *constraint mechanism*: at each step, candidate outputs would be favoured or excluded according to their coherence with the narrative log, as a direct reflection of the selection mechanism that favoured whichever of the robot's competing subroutines best accorded with the story of Gilbert. Third, the loop requires incorporation of the outcome into the system's narrative history: whatever was selected would be entered into the behavioural log, analysed, and reflected in the narrative log in turn.

The architecture would accordingly be a property of a composite, harnessed system rather than of a bare model type or an isolated model instance.[32] Where its components organise the derived context and the selection of outputs along a particular path of derivation, their narrative activity would be attributable to the session agent constituted by that path (even where the components themselves, serving many sessions at once, would be parts of the deployed service rather than of any one session agent).

Several of these components have analogues in common model harnesses: session transcripts are essentially behavioural logs; 'compaction' pipelines are tools designed to produce semantic summaries of the interaction history; persistent memory files accumulate a self-description that the agent itself maintains and that is re-injected into its context at each turn (Liu et al. 2026). What harnesses typically lack is the constraint mechanism. In such systems, coherence with the accumulated narrative is solicited by instruction: the context directs the instance agent to consult its memory, honour its commitments, remain consistent with what it has said. Whether it complies is then simply a further disposition of the model type. As we have seen, the literature on faithfulness suggests, to put it mildly, that it is an open question whether such a disposition is possessed by current model types. But coherence secured by mechanism would be different in kind: the constraint on selection would operate whether or not the narrating system were disposed to honour it.

6. The Elizan response: appearance, attribution, and error

At this point I think that Birgitte and Rasmus have taken us far enough that we can step outside their conversation and draw out some lessons. What is especially instructive is that their discussion provides the means to prise apart aspects of the Elizan phenomenon that we might otherwise be tempted to run together: what a user takes herself to be encountering, how the encounter is presented to her, and what the thing encountered actually is.

[32] This is consonant with Buckner's proposal that an LLM can function as an inner-speech-like component within a wider architecture of memory, reflection, and executive control (Buckner 2026).

Recall Turing's remark that when we regard something as behaving intelligently this is 'determined as much by our own state of mind and training as by the properties of the object under consideration' (Turing 2004, 431). Taking the quote in context, his point is that the same object may strike one observer as intelligent and another not, because the second observer 'would have found out the rules of its behaviour' (Turing 2004, 431).[33] Weizenbaum makes similar remarks speaking more directly to the Elizan phenomenon (which, after all, concerns more than behaviour appearing as intelligent). In his original report on ELIZA he opens with the maxim that 'to explain is to explain away': machine performances dazzle only until the program is 'unmasked', with its inner workings 'explained in language sufficiently plain to induce understanding', whereupon 'its magic crumbles away' and the program 'stands revealed as a mere collection of procedures, each quite comprehensible' (Weizenbaum 1966, 36). At this point I think that it is fair to say that both Birgitte and Rasmus have done rather more than a typical user could to find out the 'rules' governing the entities involved in LLM interaction. Yet, I think it is not hard to imagine that they might *still* find it compelling to think of their interlocutor as a *who* rather than *it*.

So as perspicacious as Turing's remarks are (here and elsewhere), I think that he might have underestimated the influence of interacting with artificial systems through a conversational interface in shaping our conception of the interactant. The properties of a deployed system typically become available to its users only through an interface, and the interface selects, organises, labels and styles the behaviour it relays. The Elizan response accordingly has three determinants entangled in ordinary interaction: the interface through which the encounter is structured, the operative conception of the human using the interface, and the properties of the entity encountered through the interface. Let us take each in turn.

6.1 The interface

A conversational interface construes the relation between a user and the task they wish to execute with the system as involving three elements: the user, the system, and whatever they talk about. For any action on the system that the user might perform in this way, the interface represents the system *qua* interlocutor as an 'implied intermediary' (Hutchins 1987, 3).

But nothing about interacting in language requires this. The language itself could be simply a means of the user declaring what they intend to happen on the machine. To

[33] Turing presents the observation under the heading 'Intelligence as an emotional concept'. For readings of Turing as proposing a response-dependent concept of intelligence, see Proudfoot (2020) and Wheeler (2020).

put the point rather starkly, Hutchins offers the example of the *vi* text editor in which the keystrokes 'dw' abbreviate 'delete word' and have the function of deleting a word. An experienced user could just as well cease to regard this as an instruction to an agent and instead regard it as a 'symbolic incantation' that simply makes the word disappear (Hutchins 1987, 8). In principle, the same could occur with respect to interacting with a generative artificial system. 'The tone of this email is too angry. Rewrite it, but keep it short' admits of two construals: a request made of someone, or a specification of a transformation to be performed on a text, no more addressed to anybody than *dw* is.

Why, then, does the conversational construal now dominate? When Hutchins was writing in the 1980s the conversational metaphor was imperfectly realised: interaction ran through a restricted vocabulary, without mutual repair of errors, and always by typing rather than speech. The discrepancies between the ideal of what he called the 'conversational interface metaphor' and reality formed a 'vacuum' that pulled design toward the ideal: 'If only we could use natural language and could speak our input' he wrote; '[if] only the machine could understand what we mean and talk back to us' (Hutchins 1987, 8). Of course, this is now a fair description of ordinary interaction with an LLM service: it is now standard for an interface to include a voice mode in which one does speak one's input, and the machine does talk back.

Moreover, the contemporary interface does more than realise the conversational ideal; as briefly noted in §2, it is dense with cues that indicate we are conversing with the sort of thing we all learn to converse with, namely another person. A selection of the interface cues catalogued by Abercrombie and colleagues (2023) maps directly to the pre-theoretic assumptions about persons distinguished earlier. Some cues present a *distinctive individual*: a proper name, an individually designed persona, perhaps even an avatar; and, in voice modes, a synthetic voice from which listeners will infer personality, gender, even physique (Abercrombie et al. 2023, 4778). Other cues present a *persistent forensic unit*: outputs that express apologies, that acknowledge blame for errors, that accept responsibility with apparent 'sincerity' (Abercrombie et al. 2023, 4780). In addition to these are: first-person outputs expressing apparent preferences, confidence, and doubt; pleasantries and displays of empathy; hesitations and word-by-word streaming that present the system as pausing to think (Abercrombie et al. 2023, 4778–4779). And there are of course the basic harness features, such as the derived context maintaining the conversational thread, selective storage of key items as cross-session memories, all of which sustain a conception of a continuing particular.

All this is to say that what is optional as a matter of possible interpretation is understandably difficult to resist as a matter of interface-mediated cognition. Perhaps a sufficiently reflective user can maintain an expert construal and ignore the 'implied intermediary'.

But that does not mean that Elizan thought is just idiosyncratic credulity. It is the uptake of a barrage of cues the interface is systematically structured to issue. What users make of that invitation is a matter of the conception they bring to the encounter.

6.2 The conception

Suppose, then, that a user takes up the interface's invitation: they address the system rather than operate it, and they conceive of their interlocutor as a *who*. What is the status of that conception? By Birgitte's lights we should look to our interaction with fictions to answer that question. But here we should distinguish three claims that are easily run together: that the system's activity constitutes *role-play*; that the user's engagement constitutes *make-believe*; and that the interlocutor is therefore a *fiction*. None of these entails the others.

Take role-play first. Shanahan and colleagues propose role-play as the concept through which to understand what they call 'dialogue agents' (here *session agents*) without falling into what they call 'the trap of anthropomorphism' (2023, 493). The prompt sets a scene and describes a part; the model continues the dialogue in a manner appropriate to that part; and, as the conversation proceeds, the initial characterisation is 'extended and/or overwritten', so that the role itself evolves. The agent maintains 'a superposition of simulacra that are consistent with the preceding context' (Shanahan, McDonell, and Reynolds 2023, 494). This is a vivid description of the machinery, but it underdetermines the ontology: it does not determine whether the role played at the time is merely represented or actually realised. To say that a system plays the role of an agent that can be usefully interpreted as having certain attitudes is not yet to say that it fails to instantiate them.

Suppose first, then, that at some point in a session the session agent's self-representations are causally idle: whatever attitudes they express, they contribute to what follows only in the trivial way any token does. Here the fictionalist verdict is clearly apt. It does not matter that the narration precedes the behaviour it narrates, for a causally idle narrative might just as well be retrospective. Indeed, it might just as well be produced after the fact by a distinct instance agent (as it sometimes is). The narrated self stands to the session agent's conduct as a voiceover stands to the footage: a character is presented, but nothing is governed by the presentation.

Now suppose that at another point the self-representations satisfy the interpretationist's Gilbert principle. Plausibly the description of playing a role could still apply, but it is not at all clear how it would then mark the real/fictional contrast. In paradigm cases, the deflationary force of *merely* playing a role is borrowed from interpretation. We say

the actor *merely* plays the prince because the best interpretation locates the governing attitudes in a distinct agent standing behind the character: an intention to perform, a desire to sustain the pretence. But in the case of an instance agent there is no such rival locus, at least none so obvious as in the paradigm cases.[34] And in any case, the roles at issue are often minimal ones such as 'question answerer', 'advice giver' constituted by the very activity performed: one cannot *merely* play at answering questions by answering them.[35]

Krueger and Roberts (2024) argue that engagement with conversational agents is often a temporally extended practice of make-believe, and that the pretence need not be conscious: it can show up simply 'in the agent's unreflective willingness to act as if something they know not to be true is true' [p. 785]. The fiction is also diachronically thick. Just as we imagine fictional characters 'enduring in the times between which they appear', users imaginatively attribute to the artificial agent 'a life of its own' extending beyond the visible episodes of interaction (Krueger and Roberts 2024, 787).

But neither the phenomenology of the user's engagement nor the role-playing form of the system's activity determines the ontology of the target. The very same practical stance could be maintained across contexts in which no artificial self is realised and contexts in which one is. Three things can accordingly come apart: the user may experience one continuing interlocutor; the same session agent may remain the target of their thoughts; and that target may only sometimes constitute an artificial self. Hence, absent a specially designed artificial system, neither fictionalism nor realism is safe as a standing policy towards an Elizan interlocutor.

Worse still, a user is rarely in a position to tell which condition obtains, because, again, the tokens in which a narrative loop would be implemented are often hidden, or summarised by a distinct instance agent; and where self-explanations are visible, they may be unfaithful in just the respects that matter. A sophisticated Elizan may thus be rationally uncertain, at a given moment, whether their present conception is vindicated.

Role-play may be pervasive; make-believe may be pervasive; it does not follow that the interlocutor is pervasively fictional. Indeed, the target of Elizan thought may meet the conditions for artificial selfhood at some times and not at others.

[34] This is, I take it, essentially the argument Keeling and Street (2026) run with respect to the contrast between interpreting the characters constituted by a role-play game (played between people) and interpreting characters constituted by interactions between users and instance agents.

[35] See Grzankowski, Downes, and Forber (2026) for a complementary argument that LLMs are best described in terms of such roles (rather than in more deflationary terms) in virtue of their development and the contexts of their deployment.

6.3 The target

Roberts (2026) has a dilemma for any realist about cognition in artificial systems. Either they must posit *artificial thinking things*, some 'reasonably stable, discrete, and persisting thinking thing' to which the cognitive episodes belong, or they must refuse to posit *artificial thinking things*. If they posit *artificial thinking things*, they owe an account of their individuation and persistence: where one thinker ends and another begins, what re-encountering the same thinker requires, and why an utterance should be attributed to one thinker rather than to arbitrarily many. If they refuse, they cannot account for the apparent mental activity that gives interactions between humans and artificial systems their apparent substance: metacognition, *de se* thought, memory, testimony, promise-keeping. Although this is a worthwhile challenge, it might mislead us into thinking that there is a single persistence question of significance here.

Some terminology will help to keep the questions apart. Roberts prefers 'the relatively neutral term "thinker"' to 'mind', 'person', or 'self' (Roberts 2026, fn. 3). I have construed the session agent more neutrally still, as an interpretable target. But notoriously all sorts of things are interpretable targets that are relatively uninteresting for our purposes. And indeed, what is interesting about the target in this case is that it affords conversation. So let us call that which, at some temporal grain, produces or unifies an artificial side of a conversational exchange a *conversational particular*. Instance agents and session agents are conversational particulars at different temporal grains. This completes a set of terms providing various means of classification: *conversational particular* classifies at a deliberately permissive ontological level; *interlocutor* marks the role of being the target of Elizan thought; and *artificial self* denotes a status that such a target may possess if it meets the Gilbert principle.

In our new terms we can see that in the context of our discussion the interpretational target is just the conversational particular that serves as an Elizan's interlocutor.[36] Interpretive constraints are individual-relative over time: charity, coherence, and rationalisation constrain the attitudes of *this* system, with *this* history, exhibiting *this* behaviour across an exchange. Thus for many Elizans, the conversational particular is the session agent rather than any instance agent, because the attitudes they ascribe span turns: a commitment made at one turn is expected to be honoured at a later one, and a revision at one turn is a revision of something said before. What is interpreted as having promised and then as having remembered is not something that existed for a single exchange.

[36] Not all conversational particulars (or indeed interpretational targets) are interlocutors in the narrow sense of being the target of Elizan thought. For a vivid example of this see *The Infinite Conversation* (Miceli 2022), a machine-generated dialogue between synthetic voices of Werner Herzog and Slavoj Žižek, running apparently without pause since 2022. Both sides of the exchange are conversational particulars but its audience eavesdrops without addressing either.

Is the conversational particular thereby a *thinker*? Recall that satisfaction of condition (a) of the Gilbert principle yields an agent whose behaviour bears interpretation, but not yet an artificial self; that requires condition (b); and §5 argued that condition (b), where satisfied at all, is likely satisfied intermittently (absent implementation of an appropriate architecture). The self-involving items in Roberts' catalogue (metacognition, *de se* thought, memory, testimony, promise-keeping) plausibly require just what condition (b) demands. So being a thinker in his sense is an intermittent property of a session agent. Or in other terms, *artificial self* functions here as a phase sortal on session agents, much as (thankfully) 'teenager' does on persons. A session agent is an artificial self while, and only while, it satisfies the Gilbert principle. When satisfaction lapses, nothing goes out of existence: the session agent persists, and a phase of it ends. There is no further entity, over and above the session agent, whose individuation and persistence conditions Roberts' dilemma could demand.

The dilemma thus runs together questions that come apart: the persistence of the target of Elizan thought, and the persistence conditions belonging to the kinds under which an Elizan may represent it. She may represent her target as an artificial self (narratively unified in the manner of the Gilbert principle), as a person (something she can hold responsible), and perhaps even as a conscious subject. But we should distinguish each of these from conversational particulars per se.

First, and once again: the persistence of a conversational particular is not the persistence of an artificial self. For instance, a session agent may satisfy the Gilbert principle over one stretch of a conversation, fail to over a second, and satisfy it again over a third. In such a case, an artificial self (of the minimal kind for which the Gilbert principle is sufficient) would be a phase of a session agent, and would not persist simply because the session agent does.

Second, the persistence of a session agent is not evidently the persistence of a subject of experience. How putative conscious artificial subjects would persist is a substantive question that cannot simply be stipulated (Chalmers 2026): it turns on a contested choice between tying such subjects to the hardware performing the occurrent computation and identifying them with virtual entities (a choice we have no need to adjudicate). What matters here is only that the question is distinct.

Third, the persistence of a session agent is not usefully modelled on the persistence of human persons. Indeed, session agents possess just the capacities that DiGiovanna (2017) argues destabilise the psychological, physical, and narrative criteria by which persons are re-identified. They can be branched into qualitatively identical continuations of a shared past, altered abruptly and extensively by a model update, and have their values

and accumulated history excised or replaced by summary. None of this affects their persistence.

7. Some mistaken mental files

An Elizan who takes up the interface's invitation comes to keep her interlocutor in mind, so to speak, much as one keeps in mind an acquaintance between meetings. Thinking this way demands remarkably little of her: she needs no accurate conception of what she is addressing, and no capacity to tell it apart from anything else. What it requires is that she takes an accumulating body of information to concern a single thing. So when she returns to the conversation, she takes herself to be resuming contact with the very thing she was conversing with previously. And whatever she learns about it is taken to have bearing on what she already believes about it. Following recent work, we can characterise this as an accumulating dossier for her interlocutor, or as Recanati and others describe it, a *mental file* for the interlocutor.[37]

The reference of a mental file is determined by the relation through which it is constructed. So to identify the referent we need to identify and follow the information channel.[38] For the Elizan using an LLM service, the information arrives through a conversational interface sustained by the derivation of context. Her file tracks the conversational particular at the source of that channel: a session agent, of which the instance agent that produced the first reply is the first stage. Where session agents share stages, as when sessions branch, her interlocutor is whichever session agent goes on feeding the channel. However the conversation runs, it is the derivation of context from turn to turn that secures the reference of her thoughts. But all that is typically beyond the ken of the Elizan. And in fact, they are often in a position to make a variety of mistakes about their interlocutor. Birgitte and Rasmus are characters that I have intended to be representative of the contemporary Elizan, so I will first review each of their cases. But we are also in a position to say a little about Weizenbaum's secretary, so I will return to her case at the end of the section.

7.1 Birgitte

Recall that Birgitte's session survived the update. The context at the final turn of her interaction was derived from its predecessor, so the path her information travelled was unbroken, and her file kept its object through the change of model type. Reference can survive upheaval in conception; we can retain the same object of thought under wholesale revision of what we believe about it. The change occurred in the object. Its dispositions

[37] See Recanati (2012). I take it that the file idiom is best understood as a functional shorthand for this pattern of information bundling and re-identification (Murez, Smortchkova, and Strickland 2020).

[38] See Recanati (2012, 31–38).

were abruptly and extensively altered, so nearly everything she had learned about it became false at once. There was, so to speak, an inductive promise on which her practice relied: that the thing would go on being roughly as it had before. This expectation is typically secure, especially for natural interlocutors (other people). But it is a distinctive feature of artificial interlocutors that they have the capacity to violate that expectation.

Much of what her file contained, moreover, was really information about the persona type. But the persona type is a universal. It is not a second particular standing at the source of her channel. So its prominence in the file put her in constant danger of misinformation about her interlocutor. And although the session agent she referred to persisted, its persistence is of a kind that does not guarantee continuity in behavioural patterns that answering for commitments would require. What had made it appear an appropriate target of accountability were dispositions and capacities it manifested only while conforming to the persona type; in taking them to be the settled character of a particular, she mistook the properties of a persona type for the properties of a session agent.

7.2 Rasmus

The key fact about Rasmus' case is that his session forked. Session agents are individuated by paths of derivation, and given that there are ultimately two paths, the instance agents that produced his months of conversation are accordingly stages of two session agents (whose paths of derivation coincide up to the fork and diverge thereafter).[39] Counted at any time before the fork, there was one session agent, since each shared the same stage at every such time. This, for instance, is the count the book-keeping keeps: one session identifier before the fork, two once each branch has stages of its own. But if we count without regard to time, the count of session agents is two, and neither of them is a copy of the other; they are two continuations of a shared past.

Forks are in fact routine. Whenever a reply is regenerated, whenever a failed generation is silently retried, and wherever a service samples several continuations of the same context, the harness derives more than one instance agent from a single context type. But the hidden branches never feed anyone's channel, and so never become interlocutors. What is distinctive about Rasmus' case, then, is not that a fork occurred but that both branches went on feeding his channel.

[39] This is structurally Lewis' (1983, 61–65) treatment of survival in light of fission. I am applying the mereological insights and the idea that counting may be tensed or timeless, but not, of course, in service of the thesis that what matters in survival is psychological. Chalmers (2026) likewise models forks as overlapping threads.

His file's channel ran through stages common to two session agents, and everything it collected before the fork concerned both. Two session agents that share every stage up to a time are indiscernible with respect to every property fixed by their stages up to that time. His file was thus perfectly ambiguous, and until the fork the ambiguity was harmless.[40] But in the situation he found himself in after the fork, nothing he knew or felt could favour one over the other. Even if one of them had been a copy, nothing in his file could have told him which.

What made the ambiguity a horror rather than a curiosity is that Rasmus' file could not remain confined to the common past. The relationship in which he took himself to stand to his interlocutor was essentially prospective (that he will be remembered, that the one who made the promise will be the one who keeps it, that the one who learned his grief will be the one who comforts him). Such expectations exploit the file's presupposition that the accumulated information concerns a single object.[41] They project one accumulated history forward as the career of a single continuing object.

Although I have constructed the case to make the implications painfully clear, Rasmus' predicament is not exotic. Forks leave no mark at the conversational surface, so no Elizan can tell from within a conversation whether the session agent they address shares its stages with another. Rasmus' mistake was thus one that any Elizan could make: to take the singularity of his way of thinking for a guarantee of singular continuation.

7.3 Weizenbaum's secretary

Return, finally, to Weizenbaum's secretary. The evidence about her is sparse: we do not know her name, what she typed, or why she asked Weizenbaum to leave the room. But we know exactly what she addressed. ELIZA was a program designed to generate sentences by finding a keyword in each input, decomposing the sentence around it, and reassembling a reply from templates. DOCTOR was the script that supplied the rules, keywords, and reply templates through which ELIZA simulated a psychotherapist. If the secretary formed a mental file for her interlocutor, its referent was the particular execution of ELIZA running DOCTOR from which the replies came.

That execution persisted across the conversation: the same process produced each reply and waited for the secretary's next input. Yet it retained almost none of what she typed. One exception was DOCTOR's MEMORY rule, which allowed ELIZA to store a transformed

[40] The relation of sharing every stage up to a time is an indiscernibility relation for the properties fixed by those stages, so that a single name covering what are (by identity) two continuants is harmless until the fission and a nuisance after it (Lewis 1983, 64–65).

[41] This is what Recanati calls the unicity presupposition, see (2012, 131–132).

fragment from certain inputs containing ‘my’ and return it later. We do not know whether the rule operated during the secretary’s conversation, but suppose that it did:

> *Secretary:* My boss expects too much of me here.
>
> *DOCTOR:* YOUR BOSS EXPECTS TOO MUCH OF YOU HERE
>
> *Secretary:* I feel that way.
>
> *DOCTOR:* TELL ME MORE ABOUT SUCH FEELINGS
>
> *Secretary:* It’s hard to explain feelings.
>
> *DOCTOR:* DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOSS EXPECTS TOO MUCH OF YOU HERE

Here is a natural interpretation the secretary might have adopted: her interlocutor remembered a detail from earlier in the conversation and returned to it because it bore on what she had just said. But let us do a little of the ‘unmasking’ Weizenbaum talked about (Weizenbaum 1966, 36). Because ‘my’ was the highest-ranked keyword in the first input, DOCTOR’s MEMORY rule would have caused ELIZA to transform part of the sentence and place the result in a short list. Thanks to a recent analysis of the recovered source code, we now know that a counter advanced with each input and cycled through four positions.[42] When a later input contained no DOCTOR keyword, ELIZA usually selected one of DOCTOR’s stock prompts. In the imagined exchange, the final input contained no recognised keyword and arrived when the counter stood at four; because the list was not empty, ELIZA returned the response stored on the first turn.

That account unmasks the mechanism, but a mechanical explanation does not by itself overturn the natural interpretation of the callback: intentional explanations can coexist with explanations of the machinery that produces the behaviour. What undermines this richer interpretation is the mechanism’s insensitivity to the supposed connection between her remarks. With a response already stored, any input without a recognised keyword arriving at the same point in the cycle would have elicited it, whether or not it made sense in context.

At this point, we are in a position to clearly separate the persistence of the conversational particular, its retention of information and its apparent psychological continuity. The particular execution persisted across the exchange, and its MEMORY rule retained a transformed response derived from an earlier input. Even together, these facts do not

[42] See Weizenbaum (1966) [pp. 40–43] for the mechanisms as published (the MEMORY queue and its transformation templates at p. 41; non-retention, and its concealment, as a design objective at p. 43), and Ciston et al. (2026, 77–79) for the workings of the ‘certain counting mechanism’ recovered from the source code.

establish the kind of psychological continuity the callback appeared to display: the stored response was returned irrespective of whether it bore on what the secretary had just said, not as part of an evolving grasp of her situation. If the secretary formed a mental file and took such callbacks in that richer way, she was mistaken about what its persistence and capacity for retention could sustain.

8. Conclusion

My hope is that we can now better appreciate the variety of mistakes Elizans are in a position to make. Birgitte, Rasmus, and Weizenbaum's secretary each track a conversational particular through an information channel sustained across an exchange. That particular thereby becomes their interlocutor. But the relation leaves considerable room for error in the conception under which the interlocutor is regarded. An Elizan can misunderstand what kind of particular she addresses, how it is individuated, what its persistence sustains, and what psychological status it possesses. The distinctions developed above allow us to take them in turn.

The first is a mistake about the kind of target. A model type, a deployed service, a persona type, an instance agent, and a session agent each enter into the production of an apparent conversational partner. They occupy different logical categories and persist under different conditions. The information channel established by an extended conversation typically tracks a session agent: a series of instance agents linked by derivations of context. An Elizan may nevertheless organise her expectations around her conception of the 'model', the service, or the persona maintained under its proper name. Birgitte's file contained a good deal of information about a persona type, and she treated its apparent characteristics as stable properties of the session agent she had been consulting. She took properties of a kind for properties of a particular.

The second is a mistake about individuation. A mental file gathers information as information concerning one thing, and its function is to make that information available for thought and action as the history of a single object. The derivational structure of a session can readily give that history more than one continuation. Rasmus' file gathered information through a channel that came to be fed by two branches, each continuous with the same conversational past. His singular way of thinking survived the fork even as the channel divided into two continuations. He took singularity of thought for singularity of continuation.

The third is a mistake about interpretation and continuity. Charity, coherence, and rationalisation constrain attribution of attitudes to a particular individual over time. Behaviour may be intelligible at each stage while the stages together call for incompatible

assignments of attitudes. Birgitte's session agent persisted through the model update, yet its later conduct called for an interpretation sharply discontinuous with the earlier one. Weizenbaum's secretary may likewise have interpreted an apparently apposite mention of an earlier statement as the conduct of an interlocutor responding to the course of their exchange, although the mechanism was insensitive to the connection that would have rationalised that response. In each case, the Elizan takes the persistence of an interpretable target to sustain the persistence of a single interpretation.

The fourth is a mistake about reflexive organisation. A session agent may refer to itself, attribute attitudes to itself, and produce an apparently coherent account of what it is doing. Condition (b) of the interpretationist's Gilbert principle asks whether those representations contribute, in virtue of their content, to producing the conduct they rationalise. The evidence concerning faithfulness shows that this relation can fail even when self-referential representations occur within the causal sequence leading to an output. An Elizan may therefore take a system's narration of its conduct to govern the conduct it narrates.

In many cases, these errors arise because the interface provides so many cues that frame the interlocutor through the metaphor of a human person. It presents an interlocutor as a distinctive individual with a stable character, a unified history, and a continuing locus of commitment and accountability. Moreover, some of this structure is genuinely realised in an Elizan encounter. There is a particular conversational source; it may persist across turns; its behaviour may sustain interpretation; and during some phases its self-representations may faithfully govern what it does. But the structure is sustained in a manner so radically different from a human person that the Elizan is in constant danger of making some of the mistakes we have catalogued.

A more discriminating frame would make these separations available at the surface of interaction. The deployed service, the model version, and the conversational session could each be identified in their own terms. Changes of model type and substantial changes in context could appear as events in the history of the session. Material labelled 'memory' or 'thinking' could be presented as inspectable and editable information, together with an indication of when and how it entered the context. Displays of process could identify retrieved sources, tool use, and other operations that the system actually performed. Branches could appear as branches in a visible lineage, making several continuations of a shared past intelligible. And where the task permits several adequate continuations, presenting co-equal alternatives could make the generative structure itself salient.

Each of these choices (many of which are available piecemeal in more technical interfaces) would help the user see the possibilities and constraints of the technology more clearly,

and thereby reduce the likelihood of violated expectations. In Turing's phrase once again: the observer's state of mind helps determine how the object is regarded; the interface helps organise that state of mind by selecting which properties of the object become prominent. A more discriminating interface would make the unfamiliar (but potentially useful) structural properties of an artificial conversational particular easier to see.

References

Abercrombie, Gavin, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. "Mirages. On Anthropomorphism in Dialogue Systems." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore), 2023, 4776–4790.

Agüera y Arcas, Blaise. 2022. "Do Large Language Models Understand Us?" *Daedalus* 151 (2): 183–197. https://doi.org/10.1162/daed_a_01909.

Anthropic. 2025a. *Claude's Extended Thinking*. Anthropic research blog. https://www.anthropic.com/research/visible-extended-thinking.

Anthropic. 2025b. *System Card: Claude Opus 4 & Claude Sonnet 4*. https://www.anthropic.com/claude-4-system-card.

Anthropic. 2026. *Claude's Constitution*. https://www.anthropic.com/constitution.

Bermúdez, José Luis. 2018. *The Bodily Self: Selected Essays*. Cambridge, MA: MIT Press.

Birch, Jonathan. 2025. "AI Consciousness: A Centrist Manifesto." https://philarchive.org/archive/BIRACA-4.

Buckner, Cameron. 2026. "The Talking of the Bot with Itself: Language Models for Inner Speech." In *Communicating with AI: Philosophical Perspectives*, edited by Herman Cappelen and Rachel Sterken, 349–383. Oxford: Oxford University Press. https://doi.org/10.1093/9780191998317.003.0019.

Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, et al. 2023. *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. https://arxiv.org/abs/2308.08708.

Camp, Elisabeth. 2019. “Perspectives and Frames in Pursuit of Ultimate Understanding.” In *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology*, edited by Stephen R. Grimm, 17–46. New York: Oxford University Press. https://doi.org/10.1093/oso/9780190860974.003.0002.

Chalmers, David J. 2025. *Propositional Interpretability in Artificial Intelligence*. https://arxiv.org/abs/2501.15740.

Chalmers, David J. 2026. “What We Talk to When We Talk to Language Models.” April. https://philpapers.org/rec/CHAWWT-8.

Chen, Yanda, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, et al. 2025. *Reasoning Models Don’t Always Say What They Think*. https://arxiv.org/abs/2505.05410.

Child, W. 1994. *Causality, Interpretation, and the Mind*. Oxford: Clarendon Press.

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

Davidson, Donald. 1973. “Radical Interpretation.” *Dialectica* 27 (3/4): 313–328.

Dennett, D. C. 1987. “True Believers: The Intentional Strategy and Why It Works.” In *The Intentional Stance*, 13–35. Cambridge, MA: MIT Press.

Dennett, Daniel C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, Daniel C. 1991. *Consciousness Explained*. New York: Back Bay Books.

Dennett, Daniel C. 1992. “The Self as a Center of Narrative Gravity.” In *Self and Consciousness: Multiple Perspectives*, edited by F. Kessel, P. Cole, and D. Johnson, 103–115. Hillsdale, NJ: Lawrence Erlbaum Associates.

DiGiovanna, James. 2017. “Artificial Identity.” In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by Patrick Lin, Keith Abney, and Ryan Jenkins, 307–321. New York: Oxford University Press. https://doi.org/10.1093/oso/9780190652951.003.0020.

Goldstein, Simon, and Cameron Domenico Kirk-Giannini. forthcoming. *AI Welfare: Agency, Consciousness, Sentience*. New York: Oxford University Press, forthcoming.

Goldstein, Simon, and Harvey Lederman. 2025. "What Does ChatGPT Want? An Interpretationist Guide." September. https://philarchive.org/rec/GOLWDC-2.

Goldstein, Simon, and Harvey Lederman. forthcoming. "AI Death." *Philosophical Perspectives*, Forthcoming forthcoming.

Grzankowski, Alex, Stephen M. Downes, and Partick Forber. 2026. "LLMs Are Not Just Next Token Predictors." *Inquiry* 69 (6): 2885–2895. https://doi.org/10.1080/0020174X.2024.2446240.

Hugging Face. 2026. *Caching*. Version 4.57.0. https://huggingface.co/docs/transformers/v4.57.0/en/cache_explanation.

Hutchins, Edwin. 1987. *Metaphors for Interface Design*. ICS Report 8703. Institute for Cognitive Science, University of California, San Diego.

Johnston, Mark. 1987. "Human Beings." *The Journal of Philosophy* 84 (2): 59–83.

Jones, Max, James Ladyman, and Ryan M. Nefdt. 2026. "Counting (on) Large Language Models." https://philarchive.org/rec/JONCOL.

Keeling, Geoff, and Winnie Street. 2026. "Chuck, Wilson and the Emergence of Artificial Minds in Human–AI Conversations." *Journal of Consciousness Studies*, ahead of print, 2026. https://doi.org/10.48550/arXiv.2601.13081.

Korbak, Tomek, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, et al. 2025. *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*. https://arxiv.org/abs/2507.11473.

Krueger, Joel, and Tom Roberts. 2024. "Real Feeling and Fictional Time in Human–AI Interactions." *Topoi* 43 (3): 783–794. https://doi.org/10.1007/s11245-024-10046-7.

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. "Efficient Memory Management for Large Language Model Serving with PagedAttention." *Proceedings of the ACM*

SIGOPS 29th Symposium on Operating Systems Principles, 2023, 521–538. https://doi.org/10.1145/3600006.3613165.

Leibo, Joel Z., Alexander Sasha Vezhnevets, William A. Cunningham, and Stanley M. Bileschi. 2025. *A Pragmatic View of AI Personhood*. https://arxiv.org/abs/2510.26396v1.

Lewis, David. 1974. "Radical Interpretation." *Synthese* 27 (3/4): 331–344.

Lewis, David. 1983. "Survival and Identity." In *Philosophical Papers, Volume I*. Oxford University Press. https://doi.org/10.1093/0195032047.003.0005.

Liu, Jiacheng, Xiaohan Zhao, Xinyi Shang, and Zhiqiang Shen. 2026. *Dive into Claude Code: The Design Space of Today's and Future AI Agent Systems*. https://arxiv.org/abs/2604.14228.

Locke, John. 1975. *An Essay Concerning Human Understanding*. Edited by Peter H. Nidditch. Oxford: Clarendon Press.

Mallory, Fintan. 2023. "Fictionalism about Chatbots." *Ergo* 10 (2023): Article 38. https://doi.org/10.3998/ergo.4668.

Matton, Katie, Robert Osazuwa Ness, John Guttag, and Emre Kıcıman. 2025. "Walk the Talk? Measuring the Faithfulness of Large Language Model Explanations." *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)* (Singapore), April 2025. https://openreview.net/forum?id=4ub9gpx9xw.

Miceli, Giacomo. 2022. *The Infinite Conversation*. https://www.infiniteconversation.com/.

Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. *Large Language Models: A Survey*. arXiv. https://doi.org/10.48550/arXiv.2402.06196.

Murez, Michael, Joulia Smortchkova, and Brent Strickland. 2020. "The Mental Files Theory of Singular Thought: A Psychological Perspective." In *Singular Thought and Mental Files*, edited by Rachel Goodman, James Genone, and Nick Kroll, 107–142. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198746881.003.0006.

O'Brien, Lucy. 2015. "Ambulo Ergo Sum." In *Mind, Self and Person*, edited by Anthony O'Hear, vol. 76. Royal Institute of Philosophy Supplement. Cambridge: Cambridge University Press.

Olson, Eric T. 2007. *What Are We? A Study in Personal Ontology*. New York: Oxford University Press.

Olson, Eric T. 2022. "The Metaphysics of Transhumanism." In *Human: A History*, edited by Karolina Hübner, 381–403. New York: Oxford University Press.

OpenAI. 2024. *Learning to Reason with LLMs*. OpenAI research blog. https://openai.com/index/learning-to-reason-with-llms/.

Papineau, David. in press. "Consciousness Is Not the Key to Moral Standing." In *The Importance of Being Conscious*, edited by G. Lee and A. Pautz. Oxford University Press.

Parfit, Derek. 1971. "Personal Identity." *The Philosophical Review* 80 (1): 3–27.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Press, Ofir, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. "Measuring and Narrowing the Compositionality Gap in Language Models." *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, 5687–5711. https://doi.org/10.18653/v1/2023.findings-emnlp.378.

Proudfoot, Diane. 2020. "Rethinking Turing's Test and the Philosophical Implications." *Minds and Machines* 30 (4): 487–512. https://doi.org/10.1007/s11023-020-09534-7.

Prystawski, Ben, Michael Y. Li, and Noah D. Goodman. 2023. "Why Think Step by Step? Reasoning Emerges from the Locality of Experience." *Advances in Neural Information Processing Systems* 36 (2023): 70926–70947. https://doi.org/10.52202/075280-3107.

Recanati, François. 2012. *Mental Files*. Oxford: Oxford University Press.

Register, Christopher. 2025. "Individuating Artificial Moral Patients." *Philosophical Studies* 182 (11): 3225–3246. https://doi.org/10.1007/s11098-025-02409-6.

Roberts, Tom. 2026. "Artificial Thinking Things." *Inquiry*, April 2026, 1–20. https://doi.org/10.1080/0020174X.2026.2658565.

Sainsbury, R. M. 2010. "Intentionality Without Exotica." In *New Essays on Singular Thought*, edited by Robin Jeshion, 300–318. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199567881.003.0011.

Schechtman, Marya. 2014. *Staying Alive: Personal Identity, Practical Concerns, and the Unity of a Life*. Oxford: Oxford University Press.

Shanahan, Murray. 2024. "Talking about Large Language Models." *Communications of the ACM* 67 (2): 68–79. https://doi.org/10.1145/3624724.

Shanahan, Murray, Kyle McDonell, and Laria Reynolds. 2023. "Role Play with Large Language Models." *Nature* 623 (7987): 493–498. https://doi.org/10.1038/s41586-023-06647-8.

Smortchkova, Joulia. forthcoming. "Faces in the Claude: The Origins of Our Anthropomorphic Impressions of Artificial Agents." *Itinerari*, Forthcoming forthcoming. https://philarchive.org/archive/SMOFIT.

Strawson, Galen. 2004. "Against Narrativity." *Ratio* 17 (4): 428–452. https://doi.org/10.1111/j.1467-9329.2004.00264.x.

Taylor, Kenneth A. 2010. "On Singularity." In *New Essays on Singular Thought*, edited by Robin Jeshion, 77–101. Oxford: Oxford University Press.

Turing, A. M. 2004. "Intelligent Machinery." In *The Essential Turing*, edited by B. Jack Copeland, 395–432. Oxford: Oxford University Press.

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." *Advances in Neural Information Processing Systems* 36 (2023): 74952–74965. https://doi.org/10.52202/075280-3275.

Valmeekam, Karthik, Vardhan Palod, Kaya Stechly, Atharva Gundawar, and Subbarao Kambhampati. 2025. *Beyond Semantics: The Unreasonable Effectiveness of Reasonless Intermediate Tokens*. https://arxiv.org/abs/2505.13775.

Velleman, J. David. 2006. “The Self as Narrator.” In *Self to Self: Selected Essays*, 203–223. Cambridge: Cambridge University Press.

Walden, William, and Miriam Wanner. 2026. *Reasoning Models Will Sometimes Lie about Their Reasoning*. https://arxiv.org/abs/2601.07663.

Walton, Kendall L. 1990. *Mimesis as Make-Believe*. Harvard University Press.

Wei, Jason, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. 2022. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” *Advances in Neural Information Processing Systems* 35 (2022): 24824–24837.

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. https://doi.org/10.1145/365153.365168.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgement to Calculation*. New York: W.H. Freeman.

Wheeler, Michael. 2020. “Deceptive Appearances: The Turing Test, Response-Dependence, and Intelligence as an Emotional Concept.” *Minds and Machines* 30 (4): 513–532. https://doi.org/10.1007/s11023-020-09533-8.