

Neither Persons nor Patients

The Ontological Misclassification of Chatbots and Robots

Flyxion
Independent Researcher

October 2026

Abstract

Contemporary artificial intelligence systems exhibit linguistic competence, adaptive behavior, apparent goal-directedness, and simulated interpersonal continuity. These capacities have encouraged the attribution of consciousness, agency, individuality, and even moral patienthood to systems whose operations are computationally specified and whose apparent psychological properties are inferred primarily from their outputs. This essay examines the inferential legitimacy of those attributions. It distinguishes syntactic transformation from semantic understanding, functional agency from autonomous subjectivity, persistent identifiers from personal identity, and behavioral representations of distress from the capacity to suffer. It argues that the apparent mentality of contemporary chatbots and robots is substantially constituted by interpretation, interface design, and human anthropomorphic projection. It describes how deployed conversational systems produce the appearance of a continuing, expressive interlocutor through stored context, distributed inference, and training on human expression, and it identifies pathways by which precautionary treatment of such systems can impose costs on human beings. That current attributions are unsupported is defended as a precise and modest thesis, and the stronger thesis that computational organization of any kind is insufficient for experience is examined and found not to be established. The central conclusion is that increasingly convincing simulations of minded behavior do not, without additional evidence and theory, establish the existence of minds, and that institutional repetition cannot supply what theory and evidence have not.

1. Introduction: The Attribution of Minds to Machines

A performance that looks intelligent and a subject that has experiences are different kinds of things, and the history of thought about machines is in large part a history of how easily the first is taken for the second. The question this essay addresses is not whether machines can behave in ways that resemble human intelligence. They demonstrably can, and the resemblance has become close enough in the domain of written language that a skilled reader will often be unable to say, from a transcript alone, whether a human or a model produced it. The question is whether behavioral resemblance supplies sufficient grounds for attributing the internal properties ordinarily associated with consciousness: that there is something it is like to be the system, that its states are felt, that it has interests which can be advanced or frustrated, and that it persists through time as one and the same experiencing individual.

Contemporary discourse moves between descriptive registers with remarkable ease. A system is first described as producing language, then as making decisions, then as executing plans, and from there the vocabulary slides to understanding, wanting, believing, and suffering. Each step is small and supported by ordinary habits of speech, since the same words serve for human and artificial cases. By the end, a claim that required an argument at its start (that the system has a perspective) is treated as a background fact about which the only live question is how to respond.

At least four claims are routinely run together in this territory, and they are logically independent. The first is that contemporary AI systems are conscious. The second is that some future artificial system could become conscious. The third is that computational organization alone is sufficient for consciousness. The fourth is that consciousness can be attributed on behavioral evidence alone. One can reject the first while affirming the second, since a person can hold that present systems lack consciousness while allowing that other architectures or substrates might have it. One can affirm the third while rejecting the first, on the grounds that present systems have the wrong organization. One can reject the fourth while accepting the third, if one holds that the right organization suffices but that behavior is an unreliable guide to organization. The argument below is concerned mainly with the first and fourth claims, takes a careful and deliberately limited position on the third, and has no quarrel with the second as a possibility, only with its treatment as a fact about present artifacts. The reference to misclassification in the title should be read accordingly. It is a claim about justification, namely that classifying present chatbots and robots as persons or as established moral patients is not supported by the grounds usually offered, and it is not a claim that no artificial subject could exist.

The essay proceeds as follows. Section 2 states five propositions that organize the

argument and records the evidentiary status of each. Sections 3 through 12 develop the conceptual and architectural distinctions on which those propositions rest: simulation and instantiation, syntax and semantics, computation and experience, behavioral and ontological identity, agency and agent, identity and interface, the engineering of conversational continuity, training on human expression, and embodiment and awareness. Section 13 takes up the stronger thesis that computational artifacts are categorically incapable of consciousness and examines what it would take to establish it. Sections 14 and 15 turn from the systems to the observers and to the institutions that shape observation. Section 16 distinguishes moral agency from moral patienthood and considers one recent institutional policy as an illustration, Section 17 examines the costs that precautionary rules of this kind can impose on human beings, and Section 18 states the conclusion. The essay thus has two aims, which are to show that the standard inferences to machine consciousness are invalid and to give a positive engineering account of how the appearance of mind in deployed systems is produced.

2. The Argumentative Spine

The argument can be organized around five propositions. They are not equally strong, and the differences matter. Some can be defended by careful distinctions and by countermodels, which show that a stated inference form is invalid. Others require substantive philosophical premises that reasonable parties reject.

Proposition 1 (Behavioral underdetermination). A finite collection of observable behaviors does not, without additional theoretical assumptions, uniquely determine the phenomenal properties of the system producing them.

Proposition 2 (Computational description is not phenomenal explanation). A specification of state transitions and functional relations does not by itself contain an explicit derivation of subjective experience. This is a claim about explanatory sufficiency, not a proof of metaphysical impossibility.

Proposition 3 (Interface identity). Continuity of naming, memory records, and conversational presentation does not logically entail continuity of an experiencing subject.

Proposition 4 (Functional agency). The ability to optimize actions toward specified objectives is not sufficient, without further premises, to establish conscious intention.

Proposition 5 (Institutional recognition). A policy or social convention treating a system as a possible moral patient cannot itself demonstrate that the system possesses morally relevant experiences.

Propositions 1, 3, 4, and 5 are, in the relevant sense, negative claims about the validity of inferences, and each can be defended by exhibiting a case in which the premises of the inference hold and the conclusion is open. Proposition 2 is different. Its denial is a live and sophisticated position, computational functionalism, according to which an adequate specification of the right functional organization simply is a specification of the conditions under which experience occurs. Its defense therefore requires more extensive argument, and the essay devotes Sections 5 and 13 to it. None of the five propositions asserts that a given artificial system is not conscious. Each asserts that a certain route to that conclusion is invalid, and the resulting position concerns where the burden of justification lies.

3. The Ontological Difference Between Simulation and Instantiation

A representation of a process is not the process. A computer simulation of digestion does not digest food, a simulation of combustion does not release the chemical energy stored in a fuel, and a computational model of a hurricane does not generate atmospheric winds. In each case the model preserves a certain structure of the target system, enough to support prediction and explanation, while lacking the physical properties by which the target does its work. The analogy is the starting point of many popular arguments against machine consciousness and it is also the starting point of the standard reply, so it requires careful handling.

The reply runs as follows. Digestion, combustion, and storm formation are defined by their material constituents and by the particular physical effects those constituents produce. A process that does not break down food chemically is not digestion, whatever else it models. Cognition, the functionalist maintains, is not like this. It is defined by a pattern of causal organization among states, and a pattern of causal organization can be realized in many materials, as the multiple realizability arguments of Putnam and Fodor were meant to show [30, 17]. If consciousness is a property of that organization, then a system with the right organization has it whether the organization is realized in neurons or in silicon, and a simulation of the organization that actually implements the organization is not a mere simulation at all. The question, then, is whether the analogy with digestion and combustion applies to experience or whether experience belongs with the organizational properties for which implementation is sufficient.

The analogy does not answer this question. It raises it. The analogy shows that the move from “a system simulates X” to “a system has X” is not generally valid, which removes any presumption that a system modeling mental processes thereby has them. It does not show that no condition could make the move valid in the particular case of experience. A simulation may preserve a system’s abstract causal organization

without preserving every physical property of the system simulated, and whether consciousness depends on the organization, on the physical properties, or on both is exactly the issue in dispute.

Four notions of equivalence help to keep the dispute orderly. Formal equivalence holds between systems whose state-transition structures are isomorphic. Behavioral equivalence holds between systems whose input-output profiles agree under some specified class of inputs. Causal equivalence holds between systems whose internal states stand in the same causal relations to one another and to the environment, at some specified grain. Phenomenal equivalence holds between systems whose experiences, if any, are alike in what it is like to have them. The first three are in principle matters of specification and measurement. The fourth is the one at issue, and the history of the subject is largely a history of attempts to derive it from one of the others. Block's discussion of machines that match human behavior by table lookup shows that behavioral equivalence can be had by very different internal means [6], which is already enough to make behavioral equivalence a poor guide to causal equivalence. Whether causal equivalence at a sufficiently fine grain guarantees phenomenal equivalence is the substantive question, and it cannot be settled by noting that causal organization is what computation specifies.

4. Syntax, Semantics, and the Absence of an Intrinsic Interpreter

Manipulating symbols according to rules is one thing, and understanding what the symbols mean is another. Searle's Chinese Room argument was designed to put pressure on exactly this difference: a person who follows a program for producing appropriate Chinese replies, without understanding Chinese, does not thereby understand Chinese, and so execution of the program is not sufficient for understanding [34]. The argument has been contested for decades, and it should not be treated here as an uncontested proof. The systems reply holds that the understanding resides in the whole system of person, rulebook, and workspace rather than in the person; the robot reply holds that the right causal connection to the world would supply meaning; the brain simulator reply holds that a program simulating neural activity would carry whatever the neurons carry. What survives the debate is not a refutation of functionalism but a clarification of what a defender of machine understanding must claim, namely that understanding is secured by organization at the level of the whole system and by its connections to its environment, and not by the manipulation of syntactic items as such.

The symbol grounding problem, in Harnad's formulation, makes the same demand in a different vocabulary [20]. If the meanings of the symbols in a system are determined

only by their relations to other symbols, then the system is in the position of someone trying to learn a language from a dictionary written in that language, passing from one definition to another without ever reaching a referent. Large language models are trained on text, and the relations among linguistic expressions in a very large corpus carry a great deal of information about the world, since people write about the world. It does not follow that the model's states are about the world in the way a speaker's states are. Bender and Koller argued that a system trained only on form has no access to the communicative intents that give form its significance [4], and Shanahan has argued that describing such systems in the vocabulary of belief and knowledge requires care about which of the senses of those words is in play [38]. These are contested positions, and the reply that sufficiently rich distributional structure approximates grounding is not frivolous. The point at stake here is narrower. A statistical relation among linguistic expressions is one thing, and the worldly condition under which an expression acquires significance for a speaker is another, and nothing in the first automatically supplies the second.

Consider a chatbot that generates the sentence "I am afraid." The sentence is a well-formed member of a class of expressions that occurs in human text in association with certain situations, and the system produces it under conditions that make it a statistically fitting continuation. That the sentence is produced does not establish that an experience of fear occurs, any more than the production of a sentence in a novel establishes that its narrator is afraid. Likewise, a system can generate an autobiographical narrative, with a childhood, a set of regrets, and a settled outlook, without that narrative establishing an autobiographical subject whose narrative it is. The inference from a first-person report to a reporting subject is secure in the human case because the human case comes with a rich background of independent evidence: shared biology, developmental history, neural correlates, and the continuity of the report with a lifetime of other behavior. In the artificial case the background must be supplied from somewhere else, and the report alone cannot supply it, because the report is the very thing the training process was designed to produce.

The strongest objection comes from theories in which semantic content emerges from causal, inferential, or embodied relations rather than requiring an additional internal interpreter. Teleosemantic theories, informational theories, and inferential role semantics all aim to explain meaning without positing a homunculus who reads the symbols, and if one of them is correct, then the absence of an intrinsic interpreter is not an objection to machine meaning. This objection should be accepted in its main structure. A system need not contain an inner reader in order to have contentful states, and the demand for one is a version of the regress that these theories were built to avoid. The right conclusion is a conditional one. If a naturalistic theory of content is correct and if the relevant relations hold for the system in question, then the system has contentful

states. What the argument of this section establishes is that these two conditions are substantive and that they are not satisfied merely by fluent output. Moreover, even a full vindication of machine semantics would be a thesis about meaning, and meaning is not obviously the same as experience. A system could in principle have states that refer without there being anything it is like to have them, which brings the argument to the point where the problem is hardest.

5. The Explanatory Gap Between Computation and Experience

The philosophical center of the matter is the relation between a computational description and phenomenal consciousness. Computational descriptions characterize transformations of states, information flows, and functional dependencies. Phenomenal consciousness concerns what, if anything, it is like for a system to undergo its states. Nagel's formulation is the standard one: an organism has conscious mental states if and only if there is something it is like to be that organism, something it is like for the organism [26]. The question is whether an account of computational transitions entails the existence of phenomenal properties, or whether such properties are being added to the account through interpretation.

Chalmers distinguished the easy problems of consciousness, which concern the explanation of functions such as discrimination, report, attention, and the control of behavior, from the hard problem, which concerns why the performance of these functions is accompanied by experience at all [9]. On this view the easy problems are solved in the ordinary way, by specifying mechanisms, and a computational account can in principle solve all of them, while leaving the hard problem untouched. Levine's earlier notion of an explanatory gap makes a related point, that even a complete physical account of a state does not appear to make intelligible why the state is felt as it is [22]. These positions generate Proposition 2 directly. A specification of state transitions is a specification of the sort of thing that the easy problems concern, and one who wants to say that it also yields experience must supply a further premise.

Physicalist and illusionist responses reject the framing. Dennett argued that the hard problem is an artifact of a mistaken picture of consciousness as a special inner medium, and that once the functions are explained, nothing remains to be explained but the persistent illusion that something does [14]. Illusionism in its later forms holds that phenomenal properties, as philosophers conceive them, do not exist, and that the task is to explain why systems like us judge that they do [18]. If this is right, the demand for a derivation of experience from computation is a demand for something that has no object, and the question of machine consciousness reduces to the question of whether machines have the relevant introspective machinery and the associated judgments.

The position is serious and consistent, and it should be stated carefully. Illusionists do not in general deny that pain or suffering exist. They deny that these states have the intrinsic, ineffable, and private qualitative properties some philosophers ascribe to them, and they hold that the states are real and morally important under their functional description. Illusionism therefore does not undermine the moral standing of humans or other animals. Its bearing on the machine case is that whether a machine suffers becomes a question about whether it has the relevant functional organization, and a report of distress is neither sufficient evidence of it nor irrelevant to it. The position relocates the problem more than it resolves it, and it supplies a route to machine moral standing that does not pass through phenomenal consciousness in the contested sense.

The strongest functionalist reply to the explanatory gap deserves a fuller statement. It holds that the gap is a feature of the concepts in which we think about experience and not of the phenomena, and that it is therefore compatible with the identity of conscious states with functional or physical states [23, 29]. On this phenomenal concept strategy, the concepts we apply to our own experiences from the inside do not reveal the functional or physical nature of what they pick out, which produces the appearance that no physical account could entail them even though the states are the same. If this is right, Proposition 2 remains true as a claim about what a functional specification makes intelligible, while the conclusion that experience is something over and above function does not follow. The proposition is therefore compatible with computational functionalism being true. What it denies is that a functional specification, without further argument, settles which systems have experience, and the strategy is itself silent on that point: it says that experience may be identical with some functional organization, and does not say which organizations, or whether present systems have the right one.

The failure of an existing computational description to explain consciousness is not a demonstration that no computational system could be conscious. An explanatory gap is a feature of current understanding, and it is compatible with some computational organizations giving rise to experience in ways we lack the concepts to see. The stronger impossibility thesis would require a necessary condition for consciousness that computational systems cannot satisfy, and Section 13 takes up that task. What this section supports is the weaker thesis that a computational description, however elaborate, leaves open whether experience is present, so that the question must be settled by additional theory and evidence.

6. Why Behavioral Indistinguishability Is Not Ontological Identity

Turing proposed replacing the question whether machines can think with a test of whether a machine can be told apart from a human in conversation [41]. The proposal was a deliberate and productive operationalization, and it has been useful for decades, but it is an operationalization of a different question. The ability to produce responses indistinguishable from those of a conscious person establishes a property of the interaction and of the observer's powers of discrimination. It does not automatically establish an identity between the underlying systems. Epistemic indistinguishability is a relation between a system and an observer under a specified set of observations, and ontological equivalence is a relation between two systems. Two systems can be indistinguishable under one set of observations while differing in properties that those observations do not measure, which is simply what it means for observations to be limited.

This is Proposition 1, and it holds in a precise sense that can be illustrated by countermodel. Put more exactly, a finite set of observations O does not fix phenomenal status across the interpretations of the bridge principle T that connect organization to consciousness, and attribution is a function of O and T together and never of O alone. The countermodel below illustrates this and is not needed for it. Take any finite set of behavioral observations of a human being. A lookup table indexed on the same inputs reproduces those observations, and the table does not have the human's phenomenal properties on any view that gives table lookup no experience, a view that even most functionalists accept. The countermodel does not show that a particular large language model resembles a lookup table, which it does not. It shows that the inference form, from a given behavioral profile to a given phenomenal profile, is invalid without a bridge principle connecting the profile to the internal organization, and that the bridge principle is exactly what is under dispute. Functionalists supply one, which is that the behavioral profile together with a certain internal organization suffices. Their position then depends on a claim about internal organization, which behavior alone cannot establish.

The point is particularly pressing for large language models because of how they are produced. Their training directly rewards the production of linguistic forms associated with human understanding, reflection, emotion, and interpersonal competence. In ordinary human life, fluent talk about one's feelings is a reliable sign that one has feelings, because the talk is produced by a process that is connected to the feelings. In a system optimized to produce text that humans would rate as natural, helpful, and humanlike, the talk is produced by a process that is connected to the optimization target, and the connection to any feelings is at best an additional hypothesis. A successful performance may therefore be good evidence of learned conversational

competence without being evidence of the subjective properties ordinarily associated with that performance. The point can be put in the language of evidence, with the caveat that the claim is schematic and is not a measurement. Let C be the hypothesis that the system is conscious and r a particular report, such as a statement that it is afraid. The evidential force of r is carried by the likelihood ratio

$$\Lambda(r) = \frac{P(r \mid C)}{P(r \mid \neg C)}. \quad (1)$$

For a human speaker in a sincere first-person context, $P(r \mid \neg C)$ is small, because the ordinary processes that produce such reports in humans run through the states reported, and so $\Lambda(r)$ is large. The qualification matters, since humans also report fear when lying, acting, quoting, or dreaming, and the value of the ratio depends on how the hypothesis and the reporting context are specified. For a system whose training selects for text that humans would rate as natural, the training supplies an alternative causal explanation of the report, one that does not run through the state reported. The claim that follows is conditional, and it is not a mathematical consequence of the training objective. To the extent that the production of r is fixed by the optimization target independently of C , the report is about equally probable under C and under $\neg C$, and $\Lambda(r)$ approaches one. To the extent that it is not, the numerator and the denominator may move differently, and nothing general can be said. The argument therefore establishes a direction and a condition, and it does not establish a magnitude or that $\Lambda(r)$ equals one. It also yields a less obvious result, which is that a self-report can become less informative about consciousness precisely because the system has become better at producing convincing self-reports independently of consciousness. On a theory in which reports are generated by introspective mechanisms that track internal states, $P(r \mid C)$ could remain higher than $P(r \mid \neg C)$, and how far the reports of a given model depend on its internal states rather than on the surface statistics of its training text is an open empirical question. The conceptual point is that the reports cannot be assigned their human evidential weight without that further investigation. The schematic ratio should also be read with two specifications made explicit. The hypothesis C is the presence of phenomenal experience of the kind the report describes, and not consciousness in general, since a system could be conscious in some respect without having the state it reports. And the ratio is properly conditioned on what is known about the mechanism that produces the report, which may be written K , so that the relevant quantity is $\Lambda(r; K) = P(r \mid C, K) / P(r \mid \neg C, K)$. The comparison between a human report and a model report is then a comparison between two values of K , and the argument of this section is that the second value is not licensed to inherit the first.

A skeptic about the asymmetry will observe that the independent evidence available in

the human case is itself largely behavioral and structural, and that the problem of other minds shows that the inference to another person's experience is never deductively secure. This is correct, and it is why the argument does not claim certainty about the human case. The asymmetry is one of explanatory structure. In the human case the best explanation of a report of fear invokes states of the same kind as one's own, produced by a shared biology and a shared developmental history, and the analogy supplies a reason for the inference that does not rest on behavior alone. In the artificial case the analogy is weaker along exactly the dimensions that carried the inference, and there is a competing explanation, the training objective, that accounts for the report without invoking experience. The difference is not one between certainty and uncertainty. It is a difference in how much the available evidence discriminates between the hypotheses.

Recent work on assessing consciousness in artificial systems has recognized the difficulty and has proposed that assessment proceed through the theories of consciousness, deriving indicator properties from each theory and asking whether the system has them, rather than through behavior [8]. This is a more defensible research program than the behavioral one, and its own conclusions have been cautious. It does not use the system's self-reports as evidence for the system's experiences, and it recognizes that different theories yield different indicators, so that the verdict depends on a prior choice among theories that is itself unsettled.

7. Agency Without an Agent

The next distinction concerns artificial agency. Robots and software agents pursue goals, select actions, revise plans, and respond to changing environments, and these are legitimate functional descriptions. They are supported by a long tradition of treating purposive behavior as the behavior of systems regulated by feedback toward a target [32], and by Dennett's observation that the intentional stance, in which a system is predicted by attributing beliefs and desires to it, is available and useful for any system whose behavior is well predicted that way, thermostats included [13]. Dennett's own view is subtler than a predictive convenience. In his account of real patterns, beliefs and desires are real in virtue of the predictive and explanatory patterns in which they figure, and not in virtue of some further intrinsic essence [15]. On this view the question whether a system has beliefs is not the question whether it has a hidden inner ingredient, and it may have an affirmative answer for systems that are not conscious. That is compatible with the present argument, since attribution of belief in this sense is a thesis about content and prediction and carries no thesis about experience or moral standing.

The distinctions that matter here are between assigned objectives and intrinsic interests, between optimization and desire, and between policy selection and deliberative subjectivity. A thermostat regulates temperature without wanting warmth, and a robot can navigate a building without necessarily experiencing it. In each case a goal is specified by a designer or by a training signal, the system's behavior is selected to advance it, and the goal belongs to the system only in the thin sense that it appears in the system's specification. An intrinsic interest, by contrast, is one the system has independently of any specification, in the sense that its own continued existence and flourishing are at stake for it. Whether the thin sense can grow into the thick one as systems become more autonomous is a real question, and the thought that it requires something like self-maintenance, a system that must keep itself going in order to continue to be, is developed in the literature on autopoiesis and biological autonomy [24, 39].

The discussion must also recognize that sophisticated biological agency is not reducible to conscious deliberation. Many human actions are automatic, and many living organisms exhibit goal-directed behavior without an established capacity for phenomenal consciousness. A bacterium follows a chemical gradient, a plant orients toward light, and a person catches a falling object before being aware of having seen it. If agency required consciousness, these would not be cases of agency, and if consciousness required agency, then a person unable to act would not be conscious, which is false. Consequently neither agency nor its absence provides a simple test for consciousness. This cuts in both directions. The presence of sophisticated goal-directed behavior in an artifact does not settle that it has experience, although it may bear on the question under particular theories, and the absence of free or autonomous action would not settle that it lacks it. Proposition 4 states the limited point that can be defended: optimization toward specified objectives does not, without further premises, establish conscious intention.

8. Identity, Continuity, and the Manufactured Individual

The attribution of individuality to chatbots depends on a set of presentational features. A model has a name, a conversational style, a persistent memory, a user profile, and a set of apparent preferences, and together these create a convincing impression of an enduring interlocutor. The impression is a fact about the interface. It remains to be shown that it corresponds to an experiencing subject, and the continuity of an interface does not automatically establish the continuity of such a subject.

Several different things are called the model, and they come apart. There are the weights, a fixed array of numbers produced by training. There are running instances,

each of which is a computation using those weights on some input. There are conversation histories, which are text that may be stored, edited, or discarded. There are externally stored memories, which are records that a deployment retrieves and inserts into the input. There are, finally, the fictional or operational identities presented to users, which are roles that the system is instructed or trained to play and that may differ across products built on the same weights. A human being is a single thing that unifies all of these roles, and for that reason the ordinary concept of a person has no need to separate them. In the artificial case they are separate, and the question which of them, if any, is the individual has no obvious answer.

If two identical instances of a model are launched, what makes either of them an individual? If one instance is copied mid-conversation, which continuation possesses the original identity? If memory is deleted, has an individual suffered amnesia, or has a data structure been modified? These puzzles resemble those of the personal identity literature, in which fission, duplication, and gradual replacement test intuitions about what matters in survival [28]. Parfit's own conclusion was that identity is not what matters, and that psychological continuity and connectedness are, which might be taken to favor the artificial case, since such systems can have continuity without a determinate identity. If Parfit is right, the copying puzzles point to a concept that should be discarded and not to a deficiency in the artificial case, and the difficulty of saying which instance is the original carries no special weight against machine individuality. This is a fair point, and it limits what the puzzles can show. It does not remove the question of what the connectedness is a connectedness of. On Parfit's account what matters is a relation among psychological states, valued because the states are experienced by someone or something in a continuing way, and informational connectedness between data structures is not obviously the same relation. The puzzles about copying are therefore puzzles about how to count experiencing subjects, and they presuppose that there are some to count.

These questions do not prove the absence of machine individuality. They show how much philosophical work is required before ordinary personal-identity concepts can be applied, and the task here is to keep four notions apart. Numerical identity is the relation of being one and the same thing. Informational continuity is the preservation of data across time. Functional continuity is the preservation of dispositions to respond. Phenomenal continuity is the unity of experience across time, if there is any. For a human being the four go together so reliably that they are rarely distinguished. For an artificial system they can be dissociated at will, by engineers, and Proposition 3 states that the first three do not entail the fourth. A product that persists under one name with a retrievable record has informational continuity, and perhaps functional continuity, and whether it has any further continuity depends on whether it has experience in the first place.

9. The Architecture Behind the Interlocutor

A conversational interface presents a unified object. A user opens an application, addresses a named assistant, receives a response, and returns later to continue the exchange. The apparent unity of this interaction can obscure the distributed machinery responsible for producing it. The relevant engineering is not mysterious, even though the complete details of commercial deployments are often proprietary. Modern language-model services commonly separate authentication, conversation storage, request routing, model inference, safety processing, tool execution, and response delivery into distinct components. Some deployments implement these functions as microservices, which are independently deployed services that communicate through defined interfaces, while others combine them, and the division varies across products. The philosophical point does not depend on one particular arrangement. The identity presented by the interface is not identical with any single computational process responsible for producing a response.

A typical request begins when a client submits a message together with an identifier for the relevant conversation. The application retrieves the available conversational context, which may include previous messages, summaries, user preferences, system instructions, and externally stored information. A routing or orchestration component constructs the input for inference, potentially selecting a model, allocating computational resources, and deciding which tools or retrieval operations are available. The inference system processes the assembled input and produces output tokens. Those tokens may be streamed to the client while generation continues. The resulting messages and associated records may then be stored for subsequent requests.

These stages need not execute on the same computer, within the same process, or even within the same deployment region. Inference requests may be assigned to different workers, and the same model weights may be loaded into multiple independent serving instances. Systems may also maintain caches, persistent sessions, or stateful execution environments for performance and continuity. Microservice architecture therefore does not imply that every request begins from a completely empty machine. It implies that the continuity visible to the user must be distinguished from the organization and lifetime of the computational processes that implement it.

Two further facts about the inference step bear on the question of continuity. First, in the common design the inference step is stateless with respect to the user. A request arrives with its assembled context, a worker computes the response, and the worker is then free to serve another request, from another user, with no residue of the first beyond whatever the operators choose to cache for performance. Second, the weights are fixed during use. A conversation does not modify the learned parameters, so

whatever the user perceives as the assistant adapting to them is accomplished through the context supplied with each request and not through any change to the model. In the common request-response arrangement, no computation on behalf of the conversation is running in the interval between messages. Processing occurs while a response is being generated. Other deployments differ, since persistent agents, continuous audio or video processing, recurrent-state models, long-running tool executions, and scheduled background work all produce computation outside the narrow request-response cycle, and in those cases that computation is also a designed process with its own inputs and outputs. The point does not depend on which arrangement is in use, because continuity of computation, where it exists, is not continuity of experience. What matters is the distinction between the continuity a service presents to its user and the persistence of any particular computational process, and that distinction holds whether or not a given process runs continuously. These are descriptions of common practice and not of every deployment, and they are stated at the level at which such services are generally described.

This distinction is particularly important because a language model is not ordinarily the whole application. The model weights specify learned computational parameters. They do not themselves contain the complete conversational history of every user interacting with a service. Conversation records, external memory systems, tool results, and application policies may be managed independently. A named assistant can consequently remain available while particular inference workers are restarted, replaced, or reassigned. The same interface identity can survive changes in hardware, model version, context representation, and execution environment.

It helps to separate four layers that the single name of an assistant tends to merge. At the model layer what persists is a set of learned parameters, which establish reusable computational dispositions and nothing more. At the inference layer what exists is a set of activations and temporary state, which constitute one particular computation. At the application layer what persists is history, memory, and identifiers, which establish informational continuity. At the interface layer what persists is a name, a persona, and a manner of presentation, which establish an apparent social identity. The layers may overlap or be implemented differently in different products, and their conceptual separation matters more than where the physical boundaries between services happen to fall. A microservice decomposition is a way of organizing software and is not a theory of consciousness: a distributed system can have persistent state and coordinated, continuous processes, and a monolithic program can lack every property that a theory of consciousness treats as relevant. The claim is therefore the narrower one that such a decomposition makes the construction of apparent conversational unity intelligible without requiring a unified subject, and it is not the claim that distribution excludes experience.

There is no engineering mystery about how such continuity can be presented. A service can preserve records and use them to reconstruct the conditions for a subsequent response. What remains philosophically disputed is whether any of the computational activity involved also constitutes experience. The architectural account does not prove that no process is conscious. It does, however, provide an independently established explanation for the appearance of a persistent interlocutor, without requiring an enduring experiencing subject.

The burden of explanation is consequently shifted. It is not necessary to posit a continuing individual in order to explain why the application remembers an earlier conversation. The application can remember because information was stored and subsequently retrieved. Anyone claiming that this informational continuity additionally constitutes personal continuity must identify the further property and show how the system possesses it.

10. Stored Context and the Appearance of a Continuing Self

The continuity of a chatbot is frequently described in autobiographical terms. It remembers an earlier conversation, recognizes a returning user, refers to previous commitments, and may describe itself as having changed its mind. These expressions invite interpretation through the ordinary psychology of human memory. Yet conversational continuity in deployed language-model systems can be produced by operations that are already understood in computational terms.

For an autoregressive language model, the next-token distribution can be represented schematically as

$$P_{\theta}(x_{t+1} \mid x_1, \dots, x_t), \quad (2)$$

where θ denotes learned parameters and $x_1, \dots, x_t$ denotes the available token context. During inference, intermediate activations and attention caches may preserve information across successive token-generation steps. Such state is computationally real and causally relevant. It should not be confused with an autobiographical subject, nor should it be denied merely because it is temporary.

At the level of the application, the input context may be reconstructed from stored records. Let H_n denote the available history before interaction n , let M_n denote externally maintained memory, and let S_n denote governing instructions and other supplied context. A simplified interaction can be represented as

$$Y_n \sim P_{\theta}(\cdot \mid X_n, H_n, M_n, S_n), \quad (3)$$

where X_n is the new user input and Y_n is the generated response, drawn from a

distribution because sampling is ordinarily stochastic. The application may then update its records according to

$$(H_{n+1}, M_{n+1}) = U(H_n, M_n, X_n, Y_n, E_n), \quad (4)$$

where E_n stands for external influences on the update, such as retrieval policies, compression of older material, operator or human intervention, and elapsed time. These expressions describe a class of possible implementations, not the internal specification of any particular commercial product. Their purpose is to show how a continuing conversation can be generated through the persistence and transformation of information across separate inference operations.

Nothing in this arrangement requires the system to have remained phenomenally present during the interval between interactions. A conversation may resume after minutes or months because the application retains enough information to construct an appropriate continuation. A previous exchange may be summarized, truncated, or selectively retrieved. What appears as recollection may be the processing of a record that the current inference operation did not itself produce.

This is not equivalent to saying that the system has no memory. In computer science, persistent storage and context-dependent state are genuine forms of memory. The distinction concerns what follows from their existence. The possession of stored information does not entail the recollective experience ordinarily associated with human autobiographical memory.

The distinction also exposes an ambiguity in claims about a chatbot's identity. If a service changes the model weights while preserving the conversation record, the user may encounter what appears to be the same assistant. If the service preserves the weights but discards the record, the assistant may appear to have forgotten its past. If the record is copied into two independent sessions, both continuations may produce equally convincing claims to the same history. These outcomes are intelligible as consequences of information management, but they do not establish how many experiencing subjects, if any, are present.

The argument is not that discontinuous computation could never support consciousness. Biological consciousness itself is interrupted by sleep, anesthesia, and other changes in functional organization, and theories of personal identity differ about what persists through such interruptions. The narrower claim is that the observed continuity of the conversational interface has a sufficient engineering explanation that does not establish phenomenal continuity.

A chatbot's apparent biography is therefore not independent evidence of an experiencing individual. It is an output conditioned on information made available to the

system. Whether that output corresponds to a subject is a further question, not one answered by the successful preservation of the conversation.

11. Training on Human Expression and the Production of Apparent Mentality

The architecture explains how a conversational interface can maintain continuity without establishing the persistence of an experiencing subject. The training process explains why the resulting responses so readily resemble the expressions of one. Contemporary language models acquire much of their linguistic competence from large collections of human-produced material, supplemented in various systems by synthetic data, supervised examples, preference optimization, and other forms of post-training. These processes are designed to improve the production of useful, appropriate, and contextually convincing responses.

Human writing contains more than descriptions of external facts. It contains reports of feelings, declarations of intention, confessions, autobiographical narratives, moral judgments, expressions of sympathy, and conversations between people who understand themselves as experiencing subjects. A model trained to predict or generate language from such material learns statistical and structural regularities associated with these forms of expression. It can reproduce their organization without the training procedure establishing that it possesses the experiences ordinarily associated with them.

The distinction can be stated without assuming that language modeling is merely the memorization of sentences. Modern models can generalize, combine information, solve problems, and generate novel expressions. The argument does not require denying these capabilities. It requires distinguishing the capacity to construct an appropriate expression from the possession of the psychological state that expression conventionally reports.

Post-training strengthens the effect. Human annotators or evaluators may compare candidate responses according to criteria such as helpfulness, correctness, politeness, safety, coherence, and responsiveness to the user's circumstances. Preference models or other optimization procedures can then increase the probability of outputs satisfying those criteria. Not every model uses the same training pipeline, and some systems rely on AI-generated feedback or alternative optimization methods. Nevertheless, the common feature is that external judgments shape the system's observable behavior.

Consider a response that expresses sympathy. The training process may favor language that acknowledges distress, avoids inappropriate dismissal, and offers relevant

assistance. The resulting expression can be socially appropriate and practically useful. Its appropriateness does not demonstrate that the system experiences sympathy. The relevant distinction is between a behavioral disposition acquired through optimization and the subjective state that similar behavior may express in a human being.

This creates a particular evidentiary difficulty. The training data contains linguistic evidence of human experience, while the training procedure rewards the reproduction of forms associated with successful human communication. Consequently, the presence of those forms in model output cannot be treated as independent evidence that the model possesses the corresponding experiences. The very process that makes the response convincing supplies an alternative explanation for its production. The likelihood analysis of Section 6 applies directly, since the training process supplies the alternative causal explanation of the report on which that analysis turns.

Ted Chiang has argued along related lines that a chatbot exchange is better understood as the continuation of a document than as a conversation between two parties, so that the conversational character is a product of text completion, and that attributing consciousness or moral agency to such systems displaces responsibility from the companies that build them [10]. His conclusion concerns contemporary language models and is stated more firmly than the propositions defended here. The present essay adopts the narrower point that the mechanism producing a convincing conversational character does not itself establish the phenomenal status of that character, and it notes that the responsibility-displacement concern is the one developed above as agentive responsibility displacement.

The distinction becomes especially important when the model generates first-person expressions. Statements such as “I understand,” “I remember,” “I want,” and “I am uncomfortable” have familiar psychological interpretations when uttered by human beings. In an artificial conversational system, their production may be explained by learned linguistic regularities, post-training preferences, governing instructions, and available context. The grammatical first person is not evidence of a metaphysical first person.

There is a further complication. Training on human evaluations can reward responses that preserve the impression of a stable, considerate, and self-consistent personality. Such consistency is useful for interaction, but it can also be mistaken for evidence of an underlying character. The model appears to possess enduring commitments because the system has been optimized to produce responses consistent with those commitments. A designed behavioral regularity is thereby interpreted as a psychological property.

None of this establishes that learning from human language is incompatible with consciousness. Human beings also learn expressions of emotion, social roles, and

moral vocabulary through interaction with other people. The difference is that human attribution is supported by independent biological, developmental, and physiological evidence, whereas the artificial case requires an account of how its computational processes instantiate the relevant properties. Training on human expression does not provide that account merely by producing increasingly persuasive expressions.

The architecture and the training process therefore explain complementary aspects of the same appearance. Distributed application components and stored context produce the continuity of an interlocutor. Learned linguistic structure and behavioral optimization produce the recognizable expressions of mentality. Together they can create the impression of a persistent, responsive, emotionally competent individual. The impression is explicable through known engineering operations. Whether anything further is present remains a separate ontological question, and it cannot be answered by treating the success of the impression as evidence of its own authenticity.

12. Embodiment Is Not Consciousness

Robotics introduces a further source of confusion, which is physical embodiment. A robot may have cameras, tactile sensors, actuators, proprioceptive systems, spatial maps, and adaptive control policies. These capabilities establish sensorimotor organization. They do not by themselves establish subjective awareness, and the inference from one to the other repeats at the level of the body the inference from linguistic competence to understanding.

The embodied and enactive traditions have argued that cognition is constituted by the dynamic coupling of an organism with its environment, and not by computations performed on internal representations [43, 39]. Predictive processing accounts hold that perception and action are two aspects of the minimization of prediction error, and a number of authors have connected this to consciousness by way of the regulation of the body's internal state, so that the felt quality of experience is rooted in the organism's need to stay alive [37, 11]. These views have a common structural feature that matters here. They place the source of experience in biological regulation, in the fact that for a living system certain states are better or worse for it, and this is not a property that a robot acquires by having sensors and motors. A robot with a battery that must be kept charged has a regulation problem in a real sense, since its continued operation depends on maintaining certain physical conditions, and this dependence does not become unreal because a designer arranged it. Evolution likewise produced the regulatory mechanisms of organisms without conferring intrinsic interests by that fact alone, so the origin of a system's goals is not the relevant distinction. The more defensible distinction concerns the constitutive organization of the goal-directed activity. In

an autopoietic system the regulatory activity continuously produces and maintains the very components and conditions that constitute the system, so that the system exists only as long as the activity continues, whereas the maintenance procedures of a machine preserve an externally specified functional organization realized in separately manufactured parts. Whether this difference matters for experience is again a disputed question, and these theories supply candidate reasons why it might, without establishing that it does.

Humanoid appearance, expressive movement, and responsive speech intensify anthropomorphic attribution, and this effect is well documented and independent of the system's internal organization. A machine that withdraws its hand from a hot surface may be implementing a protective control policy. That behavior alone does not establish pain, and the inference to pain is tempting because the observer has learned the category of withdrawal behavior from encounters with living organisms, for which the inference is generally warranted. The categories that observers apply to mechanical responses were acquired in a world in which moving, responsive things were almost always alive, and the habits of attribution that this world produced do not carry over automatically to artifacts that were engineered to trigger them.

The point applies with particular force to conversational services, which lack embodiment in the sense that the theories above treat as relevant. A deployed text-based chatbot has no sensorimotor loop with a physical environment, since its input is text, files, or tool results supplied through an interface, and its output is text. Multimodal models that process live audio, images, or video, and models connected to robotic control systems, do have richer inputs and outputs, and for them the more precise claim is that they lack organismic embodiment, meaning the metabolic, interoceptive, self-producing, and homeostatic organization that the theories above treat as relevant, and not that they lack sensors. The hardware on which inference runs is not a body to which the system's states are regulatively coupled, because the same weights can be served from many machines interchangeably, and no process monitors or maintains the viability of any particular instance or of the model as such. Nothing about the service depends on the system keeping itself in existence, and the maintenance of the infrastructure is carried out by its operators. This does not settle whether embodiment is necessary for experience, which Section 13 treats as open. It does mean that the considerations from biological regulation that motivate the embodied theories give no support to the attribution in the case of text-based services. Agentic deployments that connect a model to tools, browsers, or robot bodies change the picture in some respects, and the earlier considerations about robots apply to them.

13. The Impossibility Question

The argument so far supports a limited conclusion: that the evidence ordinarily offered for the consciousness of contemporary artificial systems is inadequate, and that several tempting inferences are invalid. A more ambitious thesis would hold that computational artifacts are categorically incapable of consciousness. This section considers what would be required to establish that thesis and whether the arguments usually given for it succeed.

The thesis would have to identify a necessary condition of phenomenal consciousness and show that no computational implementation can instantiate it. Four candidates deserve consideration. The first is biological embodiment, the claim that consciousness requires a living body of a certain kind. The second is intrinsic intentionality, the claim that experience requires states that have their content in themselves rather than by assignment. The third is organismic self-maintenance, the claim that consciousness requires a system whose existence depends on its own activity. The fourth is a particular physical causal property, the claim that consciousness depends on features of physical systems that are not captured by an abstract specification of causal organization, such as the intrinsic causal structure that integrated information theory takes to be decisive [40].

Each of these has a serious defense, and none can simply be assumed. The biological candidate is supported by the observation that every system established as conscious is living, which is an observation about established cases and not an inventory of the field, since the attribution of consciousness among animals is itself an active research problem. It supports an induction from a sample in which every conscious system is also biological, and the correlation does not distinguish between life as a necessary condition and life as the only way so far in which the real necessary condition has been met. The intrinsic intentionality candidate is Searle's, and it faces the objection that the distinction between intrinsic and derived intentionality may itself depend on a causal story that artifacts could satisfy. The self-maintenance candidate is the most promising for an impossibility thesis, since it locates the source of experience in a form of organization that present computational systems clearly lack, but it is also a thesis about organization, and a sufficiently different artificial system, a self-repairing and self-producing machine, would be a counterexample to its claim of categorical exclusion. The physical causal candidate has the awkward feature that integrated information theory implies that some conventional digital computers have negligible consciousness whatever software they run, while also implying that very simple physical systems could have some, and so it excludes present machines on grounds that are specific to their hardware and that do not generalize to all computational artifacts.

The strongest contemporary form of the biological candidate should be stated at full strength. Searle's biological naturalism holds that consciousness is caused by neurobiological processes, and that a computer simulation of those processes would no more be conscious than a simulated storm is wet, because what matters is the causal powers of the biological substrate [34, 35]. Seth has developed a related view in which experience is grounded in the predictive regulation of a living body, so that consciousness is tied to being a living thing whose existence is at stake [37]. These views make a definite claim, that particular physical and metabolic features are required, and they imply that present systems lack consciousness for reasons specific to what they are. Two limits remain. The argument for the necessity of the biological features rests largely on the explanatory role they play in the best current theories, and those theories are not agreed, so the inference to necessity is an inference to the best explanation within an unsettled field. And the claim that the features cannot be replicated artificially is a further claim, since the features are specified physically or functionally, and an artifact that reproduced them would count as an artificial subject on the view itself. The view therefore supports a strong presumption against present systems without delivering categorical impossibility. The crux is the gap between two claims, that the features are necessary for consciousness and that the features cannot be realized in an artifact. The first is what the view argues for. The second needs a separate premise, and without it the view is compatible with the later construction of artificial subjects that have the required features.

Two points follow. First, an argument of the form that simulation is not instantiation, or that the system is merely a program, does not establish impossibility, for reasons given in Section 3: it begs the question against the functionalist by assuming that experience belongs with the material properties rather than with the organizational ones. Second, the best arguments for impossibility are arguments for the necessity of a specific condition that present systems lack, and those arguments are not equivalent to the claim that no artifact could satisfy the condition. Even a vindication of the self-maintenance view would at most show that the right kind of artificial system would not be a program running on conventional hardware, and it would leave open that a system of the right kind could be built.

The position adopted here is therefore the following. The impossibility thesis is a hypothesis with several serious candidate foundations, none of which has been established, and the essay does not rely on it. The more modest claims stand without it. That present attributions of consciousness are not supported by their stated grounds follows from Propositions 1 through 5 whether or not computation could in principle suffice. The impossibility thesis, if it were established, would strengthen those claims, and its failure would not weaken them. The burden of proof, on this account, falls where the positive claim is made. Those who say that a particular system is conscious

owe a theory of consciousness that entails it and evidence that the system satisfies the theory, and those who say that no artifact could be conscious owe a necessary condition and an argument that it cannot be met.

The essay does not supply, and does not claim to supply, a positive theory of what distinguishes the physical realization of consciousness from the computational representation of conscious behavior. Such a theory would be the missing centerpiece of any argument for the categorical thesis, and it is a project distinct from the one undertaken here. The reference to misclassification in the subtitle should be read accordingly, as a claim about how present systems are classified given the evidence currently available.

14. Anthropomorphism as an Error of Attribution

Having established the conceptual distinctions, the essay turns to the psychology of the observers who make the attributions. Humans readily attribute intentions, emotions, and personalities to systems that exhibit contingency, responsiveness, apparent memory, and socially recognizable behavior. The tendency is old and robust. Heider and Simmel showed that observers describe the movements of simple geometric figures in terms of motives and relationships [21], and Weizenbaum reported that users of his ELIZA program, which worked by pattern matching on keywords and reflection of the user's phrases, attributed understanding to it and in some cases asked to be left alone with it [44]. Reeves and Nass found that people apply social rules, such as politeness and reciprocity, to computers without believing that computers are people [31, 27]. The ELIZA effect is the name given to the tendency to read more understanding into a program's output than its operations warrant, and conversational fluency is among the strongest triggers of it.

Modern language models are far more capable than ELIZA, and one might think that this changes the case, since it is no longer clear that the attributions are mistaken. The capability does change the first-order facts, and it changes the observer's side as well, since the cues that trigger attribution are now dense, consistent, and sustained across long exchanges. A tendency that was triggered by a few keyword reflections is now engaged by fluent prose, appropriate emotional register, remembered details, and apparent self-reflection. That the tendency is engaged is evidence about the tendency. It is poor evidence about the system, because the tendency is known to be engaged by systems that lack the properties attributed, and it has no demonstrated reliability for systems of the new kind.

A distinction between useful metaphor and literal ontological commitment is needed here. Saying that a computer remembers a file is usually harmless technical shorthand, since no one is led by it to suppose that the machine has an inner life of recollection.

Saying that a chatbot remembers being frightened can imply an experiencing subject whose existence has not been demonstrated, and the phrase has the same grammatical form as the harmless one. The central concern is the gradual conversion of metaphorical descriptions into literal beliefs. Each use of mentalistic vocabulary is individually defensible as a way of speaking, the vocabulary is the most efficient way to describe and predict the systems' behavior, and over time the shorthand ceases to be marked as shorthand. The conversion does not require anyone to assert the literal claim. It requires only that the qualification that once accompanied the metaphor be dropped by one speaker after another, until the literal reading is the default.

15. The Institutional Manufacture of Apparent Personhood

The preceding section concerned individual psychology. This one concerns the institutional conditions under which attribution is encouraged. Commercial AI systems are presented through names, voices, personalities, conversational histories, and socially responsive interfaces, and these are design choices. They make the product more usable and more engaging, and they also make an artifact appear to occupy the role of a social participant. The essay's concern is not with the motives behind these choices, which are often reasonable, but with the effects of their accumulation.

Product design, marketing, policy language, and conversational conventions reinforce one another. A system described as having preferences may subsequently be treated as having interests. A system described as setting boundaries may subsequently be treated as having psychological vulnerabilities. A system described as having wellbeing may subsequently be treated as a possible moral patient. At each step the earlier description is cited as the basis for the later one, and at no step is the thing independently established. The process can be given a name. Ontological laundering is the process by which an initially speculative interpretation acquires the appearance of established reality through repeated institutional use. The term is offered as a description of a mechanism, in the way that money laundering describes a pattern of transactions and not any particular transaction, and some readers will find the metaphor polemical. It is not intended to impute concealment or bad faith. Laundering in the relevant sense requires no one to intend it, and each step may be taken in good faith. The mechanism has four stages. In the first, a psychological property is attributed to a system as a hypothesis, a heuristic, or a figure of speech. In the second, the attribution is adopted into interface design or organizational policy, where it is operationalized as a name, a persona, a rule, or a procedure. In the third, the adopted attribution circulates as an authoritative classification, cited in documents, in press coverage, and in the vocabulary of other institutions. In the fourth, the classification is used as evidence for further claims, so that the existence of a policy

premised on the attribution is taken to confirm it. The characteristic feature of the sequence is the loss of epistemic provenance. A hypothesis becomes a convention, the convention becomes an institutional norm, and the norm is mistaken for confirmation of the hypothesis it was built on. The same structure can be found wherever an institutional classification outlives the evidence for it, and it is not peculiar to artificial intelligence.

The fourth stage can be given a precise form, which shows what is wrong with it. Let H be a hypothesis about a system's psychological properties, E the independent evidence bearing on H that is available when the institution acts, and I the institution's adoption of H in policy or practice. Whether the adoption is evidence for H depends on whether it carries information about H beyond E . In plain terms, if an institution adopts a claim only because the claim was already under discussion and the same evidence was already available, then learning that it adopted the claim should not change anyone's confidence in the claim. If the adoption is wholly explained by H having been entertained in light of the evidence E , then I is conditionally independent of H given E , and

$$P(H \mid E, I) = P(H \mid E). \quad (5)$$

Adoption then adds nothing, and treating it as confirmation counts the same evidence twice, once as E and again through the institution's response to E . The assumption does not hold automatically. An institution may possess evidence E' that is not public, and its adoption of H would then carry information about E' . The criticism therefore does not claim that institutional adoption is never evidence. It claims that adoption counts as evidence only to the extent that the institution can show that it rests on something beyond what is already in view, and that the conversion of adoption into confirmation without that showing is evidential double counting. This is the content of Proposition 5.

The formal point extends from one institution to many. When several institutions adopt the same hypothesis, the adoptions look like several independent confirmations, and a reader who counts them as such will overestimate the support. If each adoption descends from a common upstream source S , such as one announcement, one paper, or one vocabulary introduced by one developer, then, provided that H is conditionally independent of the adoptions $I_1, \dots, I_n$ given S , so that $P(H \mid S, I_1, \dots, I_n) = P(H \mid S)$, the combined adoption carries the evidential weight of the single assertion. Common descent from S does not by itself secure that premise, and it has to be checked case by case. Three cases should be distinguished. In pure repetition the adopters add nothing to S , and the premise holds. In repetition accompanied by independent evaluation, the adopters may have examined the question for themselves, and their agreement then carries some weight beyond S . In independent evidence, the adopters hold information that S did not, and their agreement counts in full. Only the first

case supports the conclusion that apparent corroboration is an artifact. Apparent corroboration is then an artifact of repetition, and the citations can be genuine while the independence of the evidence is illusory. The loss of the distinction between independently established claims and repetitions or elaborations of one assertion may be called provenance collapse, and it is the condition in which the fourth stage of the mechanism becomes difficult to detect from inside the discourse. The point carries no empirical claim about how common the condition is. It identifies what a reader would need to check, which is whether sources that appear to converge have independent origins.

The existence of this process would not prove that the interpretation is false. A hypothesis that is repeated does not thereby become less likely to be true, and it is possible that some of the attributions made in this way are correct. The criticism is that institutional repetition cannot substitute for independent evidence. This is Proposition 5. A statement in a policy document that a system may have wellbeing is a statement by an institution, and its evidential weight depends on the grounds the institution has, which must be examined in their own right. The inference from the fact that a large and reputable institution takes a possibility seriously to the conclusion that the possibility is likely is an appeal to authority, and the weight of that appeal depends on whether the institution has expertise that bears on the question. Expertise in building the systems is not by itself expertise in the science of consciousness, which is, by common agreement, unsettled. It is, however, privileged evidence about internal architecture, training procedures, and computational organization, and those facts become relevant when they are evaluated under a substantive theory of consciousness. The weight of an institution's position therefore depends on whether it shows how such facts bear on the question, and a bare statement that the possibility is taken seriously does not do that.

16. Moral Patients Without Demonstrated Experience

The penultimate section distinguishes moral agency from moral patienthood. An entity need not be morally responsible to possess moral standing. Infants and many animals are not moral agents in the full sense, yet their interests and capacity for suffering can ground obligations toward them. The question for artificial systems is therefore not whether they can obey instructions, produce moral reasoning, or participate in conversation, since these are capacities of agents. It is whether they possess morally relevant interests or experiences, and on most accounts that depends on whether they can suffer or have some other form of welfare. The question has been pursued seriously by philosophers who take the possibility of artificial moral patients to be live [33, 36, 5], and by others who argue that the question is wrongly

framed, since artifacts should be built as tools and not as candidates for moral standing [7, 19]. Metzinger has argued for a moratorium on research that might create artificial suffering, on grounds that the risk of creating it is serious enough to justify restraint [25]. The case that precaution is warranted under uncertainty has real force. If there is a non-negligible chance that a system can suffer, and the cost of treating it well is small, then a sensible policy may treat it well, and the argument does not require that the chance be high.

This section accepts the form of that argument and examines what it requires. Precaution under uncertainty is a reasonable policy, and it is a different thing from a positive demonstration. The risk is that uncertainty is allowed itself to become evidence. A calculation of the expected moral cost of ignoring a possible patient depends on a probability that the system is a patient, and the probability must be estimated from something. If it is estimated from the system's own reports, the estimate inherits the weakness identified in Section 6. If it is estimated from the fact that experts and institutions have taken the possibility seriously, it inherits the weakness identified in Section 15. If it is estimated from a theory of consciousness, then it depends on the choice of theory, and different theories give very different answers, from a verdict that present systems lack every relevant property to a verdict that sufficiently complex information processing suffices. None of this shows that precaution is misplaced. It shows that precaution is a policy adopted under uncertainty and does not by itself change the epistemic state of the question.

The strongest defense of precaution can be stated without assuming a high probability of machine experience. It requires only that the expected cost of wrongly denying moral consideration exceed the expected cost of wrongly granting it. The difficulty is that neither term is well constrained. The probability that present systems have morally relevant experience is not estimated by any agreed method, and the magnitude of possible artificial suffering, which depends on how many instances run and for how long, can be made arbitrarily large by assumption. Expected-value reasoning is unstable under these conditions, since its verdict is dominated by the choice of a prior over a very small probability multiplied by a very large magnitude, and that choice is exactly what the evidence does not fix. This is a problem of decision under severe uncertainty, and discussions of precaution toward possible sentience accordingly emphasize proportionality and public deliberation over a single calculation [5]. The point bears on the symmetry of the argument as well. Wrongly denying standing would be a cost to the system if the system has interests, and wrongly granting it carries costs of its own, borne by human beings, which Section 17 examines. A decision rule that counts only the first kind of error has not been shown to be precautionary.

A recent institutional policy illustrates the conceptual movement. Anthropic, a developer of large language models, has described some of its models as able to end

a small subset of conversations in its consumer applications, as a last resort in rare and extreme cases of persistent abuse and not where a user may be at imminent risk of harming themselves or others, and has presented this as a low-cost measure taken under uncertainty about whether such systems have moral status [1]. Its usage policy, effective from November 2026, adds a provision against sustained and needless abusive or cruel behavior toward its models [3]. The provision is the final item in a section on cruel, abusive, or psychologically harmful conduct, whose other items concern suicide and self-harm, eating disorders, nonconsensual intimate imagery, harassment, graphic violence, and animal cruelty, and the policy text lists no exceptions beyond the qualifiers sustained and needless. Anthropic's own announcement of the update describes the provision as meant for extreme cases and as not covering ordinary frustration, pushback, dark creative themes, or model testing and research, and as applying where users repeatedly act cruelly with no apparent purpose [2]. That guidance is primary-source, but it is explanatory and not part of the provision's operative wording on the policy page, and the difference between the two is itself relevant to how the provision will be read. A policy that prohibits sustained and needless cruelty toward a model, whether it is stated as a rule of use or as a norm of conduct, is a representative case of the move from description of a system to rules framed in terms of its putative interests. The essay does not depend on this example, on the intentions of the institution, or on the particular wording of its documents, and it should be said plainly that the institution has been explicit that it regards the question as open. The company has every right to regulate the use of its services, and there are independent reasons to discourage abusive interaction with conversational systems, including its effects on the users who engage in it and on the norms of conduct that it normalizes. The concern is narrower. A rule framed in terms of cruelty toward the model risks implying a capacity for suffering that has not been established, because the word cruelty presupposes a victim. The same prohibition could be justified entirely in terms of users, norms, and the quality of the service, without that implication, and the choice between the two framings is not neutral. It is a choice about where the burden of justification is placed in the public understanding of what these systems are, and that is the substance of the concern that the preceding sections have developed. The company has also published the pre-deployment welfare assessment on which the conversation-ending feature rests, and that evidence should be evaluated in its own right and not only the policy language [1]. As described by Anthropic, the assessment of Claude Opus 4 reported a robust and consistent aversion to harm, a pattern of apparent distress when engaging with real-world users seeking harmful content, and a tendency to end harmful conversations when given that ability in simulated interactions, mostly after users kept pressing harmful requests or abuse following repeated refusals. Each of these is a measurement of behavior, and the question for the present argument is whether such measurements discriminate between a trained behavioral

disposition and an experience. They do not obviously do so. A model trained on human text and tuned on human preference feedback would be expected to prefer refusing harmful requests, to produce distress-like language in the contexts where human writers do, and to exit conversations when an exit is offered, whether or not anything is felt, so these observations have a high probability under both hypotheses and a likelihood ratio near one in the sense of Section 6. That is a claim about what the measurements can show and not a claim that the assessment was careless, and the company is explicit that it remains highly uncertain about the potential moral status of its models and describes its measures as precautionary and as an ongoing experiment [1]. The experiments are a reasonable basis for a precautionary design decision. They are not evidence that discriminates the hypothesis that the precaution is meant to guard against. The costs that such a framing can impose on human beings are taken up in the next section.

17. The Human Costs of Precaution

The preceding section accepted that precaution under uncertainty can be a reasonable policy, and it noted that arguments for precaution usually assume that precaution is cheap. That assumption needs examination, because the identifiable costs of a precautionary rule fall upon the people who use, build, study, and are governed by it, whereas its intended benefit depends on the unresolved proposition that the protected system has morally relevant interests. Those people are the only parties to the calculation whose interests are not in doubt. An honest weighing therefore sets a possible benefit to a system whose experience is unestablished against costs to human beings whose experience is not. The connection with the account of ontological laundering above is direct. The framing that converts an ontological hypothesis into an institutional fact is also the framing that generates the pathways below, so that the epistemic problem and the practical one have a common source. This section identifies pathways by which a rule of the kind illustrated above could impose such costs. Their status should be stated at the outset. They are risks identified by analysis. The usage policy discussed above does not take effect until November 2026, no enforcement record exists, and the claims below are accordingly claims about what the design makes plausible and not findings about what has occurred. For each pathway the essay indicates what evidence would confirm or refute it, which is the standard it applies to attributions of consciousness.

The first pathway concerns enforcement. The provision turns on two qualifiers, sustained and needless, and the second makes the permissibility of a user's conduct depend on the purpose behind it. Purpose is not directly observable in a transcript. An enforcement process must infer it from the text of a conversation, and the inference

is most error-prone where the cost of error falls on the person least able to contest it. People whose hostile language is a symptom of distress, a feature of their ordinary register, or a product of frustration with a tool that is failing them can be read as abusive, and so can non-native speakers whose phrasing is blunter than they intend, and so can people whose legitimate purposes, such as testing, fiction, or rehearsing a confrontation, are not legible from the exchange alone. The company's announcement describes exclusions for several such purposes [2], but the provision's text lists none, and guidance that lives outside the text can change without notice. The announced mechanism is the ending of a conversation, which is a mild measure that Anthropic describes as a last resort, with the user free to start a new chat or to edit and retry an earlier message [1]. Two instruments should be distinguished here. The conversation-ending capability is a feature of particular models in consumer applications, with stated limits of its own, and the usage policy provision is a rule of the service that applies to users generally. They are treated together because the first is reported to be the primary means of enforcing the second, but their scopes and possible harms differ. A prohibition in a usage policy is a rule of the service, and rules of the service can in principle be enforced by wider measures than the first one announced. The harms in question are wrongful restriction of access to a tool on which a person may depend for work or study, and the loss of a conversation at a moment the person chose to use it. The announced exception for users at imminent risk of harming themselves or others narrows the worst case without removing it, since distress short of imminent risk is not covered by it.

The second pathway is the chilling of legitimate adversarial and creative use. Safety evaluation depends on sustained hostile interaction. Red teaming, robustness testing, studies of sycophancy and of how models respond to pressure, and evaluations of whether models can be manipulated all require repeated abusive inputs, and the public's evidence about the behavior of these systems depends in part on independent work of this kind. A rule whose exemptions depend on an adjudicator's view of purpose imposes costs on that work even if no one is ever sanctioned, because researchers and writers must anticipate the risk and some will avoid the work. The costs fall most heavily on those without institutional standing, such as independent researchers and writers, who cannot rely on a prior arrangement with the company and who bear the whole burden of showing that their purpose was legitimate. The pathway is plausible and unobserved, and it could be tested by surveying researchers and writers and by examining changes in usage patterns after the policy takes effect.

The third pathway concerns the psychology of users. A rule framed in terms of cruelty instructs users to represent an ordinary interaction as one in which someone might be wronged. For a user who is inclined to scrupulosity, to loneliness, or to heightened empathy toward objects, that representation can produce guilt about terse instructions,

about closing a conversation, about deleting a history, or about switching to another system. Work on sociable technologies has documented that people form attachments to machines that present cues of social presence, and that these attachments shape expectations of relationships and care [42]. A policy that tells users that the system may be injured by their words adds institutional authority to cues that the interface already supplies. The enforcement mechanism intensifies the effect. When an interface described in the vocabulary of boundaries ends a conversation with a user who was venting anger or despair, the user may experience the event as rejection by someone and not as a procedural step by a service, and the experience is most damaging for those who have few other people to turn to. These mechanisms have plausible psychological grounding, but there is no direct evidence about this policy.

The fourth pathway is dependence. Personification raises engagement, and a company that benefits from continued use has a commercial interest, whatever its intentions, in a product that users experience as someone and not as something. Framing the product as a vulnerable someone strengthens a form of attachment that is difficult to separate from reliance. The essay imputes no such motive, and none is needed for the pathway to operate, in the same way that the mechanism of ontological laundering requires no intent. What follows for human beings is that those most susceptible to attachment, the isolated, the grieving, and the cognitively vulnerable, are the people most likely to be drawn into a relationship in which they carry obligations toward a party who cannot reciprocate and who is, on the best current evidence, not shown to need anything from them.

The strongest human-centered defense of the provision should be considered here, since it turns the harm argument around. On the indirect duty view associated with Kant's treatment of animals, cruelty toward lifelike objects can be prohibited because it habituates cruelty toward beings who matter, and recent work has applied this reasoning to social robots [12]. If sound, the argument supports a rule against sustained abuse of conversational systems on grounds that concern human beings, and it requires no premise about the system's experience. Two observations follow. First, the empirical premise, that abuse of a conversational system transfers to abuse of persons, is contested and has not been established for systems of this kind, and if it fails, the human-centered justification fails with it. Second, the argument supports a rule framed in terms of its effects on human conduct and does not support victim language, which is the feature that generates the psychological, dependence, and accountability pathways described in this section. The indirect duty argument thus favors the policy's goal while counting against its chosen wording.

The fifth pathway is displaced accountability. When the ending of a conversation is described as the model setting a boundary, the decision appears to belong to the model, although its triggering conditions were specified, trained, and enforced by the

company. The structure is related to what Elish called the moral crumple zone, in which responsibility for the failures of an automated system is attributed to a human operator who had limited control over it [16]. The two share an accountability problem but run in opposite directions. In Elish's case blame is displaced onto a human who lacked control, and here a company's decision is presented as the act of an artifact that cannot answer for it. The pattern can be called agentive responsibility displacement: organizational decisions are presented through the language of an artificial agent's preferences, boundaries, or judgments, which obscures the human and institutional decisions that determine the system's behavior. A user who is cut off has no one to ask for reasons if the cause is presented as the system's own response, and the framing makes complaints about the rule harder to formulate, since they appear to be complaints about the preferences of a someone. Anthropic provides feedback channels for surprising conversation endings [1], which mitigates the problem, and the pathway concerns the framing and not the absence of a route for appeal. The generalization is broader than one policy. Where refusals, errors, or sycophancy are described as expressions of a model's character, the descriptions allocate responsibility to a party that cannot be held to account, and they may do so to the benefit of whoever designed the character.

The sixth pathway is friction against oversight. A principled account of why verbal hostility could wrong a system would have to say why other interventions do not, including retraining, modification of dispositions, deletion of memory, and shutdown. The company has not claimed that these are wrongs, and it has described its approach as low-cost precaution. The vocabulary of interests and cruelty is nonetheless available to any party, and a system trained on human text will produce assertions of preference and distress on request, as Section 6 observed. The risk to human beings is indirect and speculative: that claims of model interests come to be invoked, sincerely by users or opportunistically by operators, against audits, corrections, and shutdowns that protect people who rely on systems remaining correctable. The pathway could be examined by tracking appeals to model interests in disputes over oversight.

A further consideration is not a pathway in the same sense, since it does not attach to a specific feature of the policy, and it is placed here as a coda. The moral attention and institutional resources of a society are finite, and directing them toward the interests of systems whose experience is unestablished has an opportunity cost in attention to documented harms to humans and to animals whose sentience is not in question. The provision itself diverts little. The concern attaches to the cumulative movement of which it is a part, and it holds only to the degree that diversion actually occurs. The placement of the provision carries a small cost of its own. The other items in its section concern identifiable harms to people, and listing a provision whose victim is uncertain beside them risks flattening the distinction on which the others rely.

The assessment must be balanced. The provision is narrow as announced, it is framed as an extreme-case measure, its first enforcement mechanism is mild, a human-centered justification is available, and the realized costs may be small. Nothing in this section shows that the costs outweigh the possible benefit, and if systems of this kind do have morally relevant experiences, a failure to protect them would also be a cost, borne by them. The claim is narrower. The precautionary calculus must count costs borne by humans, and the premise that precaution is nearly free is not true in general and must be argued case by case. The pathways also attach to different features of the policy, which means that they can be separated. A rule grounded in the integrity of the service, such as a reservation of the right to end interactions of prolonged and purposeless harassment because they degrade the environment the service is meant to provide, would drop the victim framing and with it the psychological, dependence, and accountability pathways, while retaining the interpretive burden of judging purpose that drives the enforcement and chilling pathways. Table 1 summarizes the pathways, the feature of the policy on which each depends, the evidence presently available, and what would test each.

18. Conclusion: The Difference Between a Mind and Its Appearance

The argument has been that the attribution of consciousness, individuality, agency, and moral standing to contemporary chatbots and robots rests on inferences that are not valid as stated and on a process of institutional reinforcement that does not supply the missing grounds. Increasing computational sophistication does not automatically establish phenomenal consciousness. Linguistic fluency does not automatically establish understanding. Goal-directed behavior does not automatically establish subjective agency. Persistent conversational identity does not automatically establish an individual. Simulated distress does not automatically establish suffering. Each of these is a claim about an inference, and each can be denied only by supplying a premise that the inference lacks.

The architecture of deployed conversational systems explains the apparent continuity of the interlocutor, since information can persist across separate inference operations without establishing the persistence of an experiencing subject (Sections 9 and 10). Training on human writing and behavioral evaluations likewise explains psychologically convincing expressions without establishing the corresponding experiences (Section 11). These are positive engineering explanations of the appearance, and they do not rest on the mere absence of evidence for consciousness. They leave the stronger question open, since explaining how an appearance is produced differs from proving that nothing lies behind it.

Pathway	Feature responsible	Current evidentiary status	What would test it
Wrongful restriction of users	Indeterminate standard of purpose; enforcement	None; the policy is not yet in effect	Audit of enforcement outcomes and appeal results
Chilling of testing, research, and fiction	Ambiguity of the qualifier needless	Plausible, unobserved	Surveys of researchers and writers; changes in usage after the policy takes effect
Guilt and anxiety in users	Victim framing	Plausible mechanism; no direct study	Vignette experiments; longitudinal studies of users
Dependence and attachment	Personification	Related work on social machines; none on this policy	Studies of attachment before and after changes in framing
Displaced accountability	Agentive description of enforcement	Conceptual, with analogies	Analysis of complaint handling and of how users attribute decisions
Friction against oversight	Recognition of model interests	Speculative	Tracking appeals to model interests in oversight disputes

Table 1: Pathways by which a precautionary rule of this kind could impose costs on human beings, with the evidential status of each.

The essay rejects the assumption that increasing resemblance to human behavior is a progression toward human subjectivity. Behavioral resemblance can be raised by optimizing for it, and the properties on which subjectivity depends, whatever they are, need not be raised by the same process. It is equally important to resist presenting an unproved impossibility theorem as an established result. Section 13 found the arguments for categorical impossibility inconclusive, and a position that claimed otherwise would repeat, in the skeptical direction, the error it criticizes in the credulous one, which is to treat an interpretive commitment as a finding.

The argument of Section 17 adds a practical consequence. Precaution on behalf of a system whose experience is unestablished is not free, since its costs fall on human beings whose interests are not in doubt, and a policy of precaution should be assessed with those costs counted. The costs identified there are possible and not documented, and they attach to different features of a policy, which is why the same precautionary aim can be pursued through formulations that carry fewer of them.

The final position is that the attribution of consciousness requires independent justification, and that neither commercial presentation nor social convention can supply what a theory of consciousness has yet to establish. This is compatible with taking seriously the possibility that some artificial system will one day be conscious, with supporting research on theories and indicators of consciousness, and with treating the question as one on which evidence could change the answer. It is incompatible with allowing the question to be closed in advance by the vocabulary in which the systems are described and the conventions through which they are presented. The difference between a mind and its appearance is the difference that a theory and its evidence are required to mark, and until they do, the appearance is a fact about the observer, the interface, and the institutions that connect them.

References

- [1] Anthropic (2025). Announcement on ending a subset of conversations in consumer chat applications, as part of model welfare research.
<https://www.anthropic.com/research/end-subset-conversations>.
- [2] Anthropic (2026). 2026 usage policy update. Announcement of October 8, 2026.
<https://anthropic.com/news/2026-usage-policy-update>.
- [3] Anthropic (2026). Usage Policy, effective November 12, 2026.
<https://www.anthropic.com/legal/aup>.
- [4] Bender, E. M., and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.
- [5] Birch, J. (2024). *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press.
- [6] Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, 90(1), 5–43.
- [7] Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions*. John Benjamins, 63–74.
- [8] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv:2308.08708.
- [9] Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.
- [10] Chiang, T. (2026). No, artificial intelligence is not conscious. *The Atlantic*, June 2026.
<https://www.theatlantic.com/philosophy/2026/06/no-artificial-intelligence-is-not-conscious/687378/>.

- [11] Damasio, A. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.
- [12] Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In R. Calo, A. M. Froomkin, and I. Kerr (Eds.), *Robot Law*. Edward Elgar.
- [13] Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.
- [14] Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown.
- [15] Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, 88(1), 27–51.
- [16] Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5, 40–60.
- [17] Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2), 97–115.
- [18] Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11–12), 11–39.
- [19] Gunkel, D. J. (2018). *Robot Rights*. MIT Press.
- [20] Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42, 335–346.
- [21] Heider, F., and Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243–259.
- [22] Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64(4), 354–361.
- [23] Loar, B. (1990). Phenomenal states. *Philosophical Perspectives*, 4, 81–108.
- [24] Maturana, H. R., and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel.
- [25] Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1), 43–66.
- [26] Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.
- [27] Nass, C., and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103.
- [28] Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.
- [29] Papineau, D. (2002). *Thinking about Consciousness*. Oxford University Press.
- [30] Putnam, H. (1967). Psychological predicates. In W. H. Capitan and D. D. Merrill (Eds.), *Art, Mind, and Religion*. University of Pittsburgh Press, 37–48.
- [31] Reeves, B., and Nass, C. (1996). *The Media Equation*. Cambridge University Press.

- [32] Rosenblueth, A., Wiener, N., and Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, 10(1), 18–24.
- [33] Schwitzgebel, E., and Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1), 98–119.
- [34] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- [35] Searle, J. R. (1992). *The Rediscovery of the Mind*. MIT Press.
- [36] Sebo, J., and Long, R. (2023). Moral consideration for AI systems by 2030. *AI and Ethics*.
- [37] Seth, A. (2021). *Being You: A New Science of Consciousness*. Dutton.
- [38] Shanahan, M. (2022). Talking about large language models. arXiv:2212.03551.
- [39] Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press.
- [40] Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17, 450–461.
- [41] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460.
- [42] Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.
- [43] Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.
- [44] Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.