# ‘Incited and Inclined but Not Impelled’

## Distinguishing Productivity from Creativity in Artificial Intelligence

**Article Under Peer Review**


## Abstract

Discussions about the encroachment of Large Language Models (LLMs) on the domain of human creativity often conflate *productivity* with *creativity*; they conflate the ability to *generate* new structures according to rules or other systems (productivity) and the ability to *use* one’s productive capacity without fixed reliance on identifiable stimuli while nevertheless producing structures appropriate to the situation (creativity). For a behavior to be “creative” is to use the productivity of a capacity like language in ways that are stimulus-free, frequently novel, yet appropriate to situations. This paper argues that, because of this conflation, improvements in the productive capacities of LLMs have been mistaken for improvements in this richer notion of creativity. Although LLMs – underpinning both chatbots and agentic systems – exhibit remarkable productive capacities, their relation to their environment does not exhibit this “creative” aspect. Where the appropriateness of human linguistic behavior does not appear governed by a causal relationship between the individual and their environment, LLMs exhibit a functional appropriateness that sees their productive capacities tethered to identifiable stimuli in the local environment. They are, in the Cartesian sense, *impelled* to act, but not *inclined*, contrasting with the human who is ‘incited and inclined but not impelled.’


**Keywords**: Artificial Intelligence; Creative Aspect of Language Use; Generative Linguistics; Human Creativity; Free Will; Machine Productivity

## 1. Introduction

Artificial Intelligence (AI), today in its most prominent form of Large Language Models (LLMs), is evaluated by the conformance of a system's outputs with human judgments. Indeed, the improving accuracy of LLMs on any number of benchmarks may be understood as the result of sharpening the distribution of data from which a model samples generation-after-generation. This sharpening occurs in reference to human judgments, such that models of frontier generations tend to converge more effectively with human judgments while in use (see, Quattrociocchi et al., 2025). The *novelty* of an LLM's outputs – content that is, to some degree, indeterminate – results from this sharpened distribution, such that they are not reproductions of the model's training data (most of the time[1]), but nevertheless *useful*.

The nature of these improvements are improvements in the *productivity* of AI systems, in the technical sense of "the power to combine a given set of elements in accord with a given system or set of rules in order to get a 'new' structure" (McDonough, 1993, 39). LLMs, trained on vast training datasets comprised of human-generated and synthetic data, acquire productive capacities which can then be recruited by humans.

Debates over the "creativity" of LLMs in relation to the flesh-and-blood kind often conflate *productivity* with *creativity*; that is, what it means for a human to be "creative" is often judged in reference to the ability to produce new structures (typically in specialized domains) – linguistic expressions, mathematical proofs, etc. That these new structures increasingly conform with human judgments casts a spell over such debates, as the coupling of novelty with appropriateness seems to diminish the uniqueness of human creativity.

This paper revives an old, richer notion of creativity defined as *the stimulus-free and unbounded expression of appropriate thought*. The arguments herein are motivated by a Cartesian observation about human and machine behavior, summarized by Chomsky:

> Descartes argued that the creative use of language marked a distinction between human beings and other animals, and between human beings and machines. A machine may be impelled to act in a certain way, but it cannot be inclined; with human beings, it is often the reverse (Chomsky, 2017, 1).

The "creative" use of language does not refer merely to language's *infinite productivity* - the capacity to produce and understand an unbounded range of structured expressions. It refers to the *use* of language in arbitrary circumstances, such that language is exercised independently of identifiable stimuli in the local environment, yet its exercise is *appropriate* to the situation as judged by other human participants. LLMs do not exhibit this "creative" aspect of behavior, and neglecting the original Cartesian insight, adapted to a contemporary context, risks a continued conflation of improvements in their productive capacities with human behavior that is "incited and inclined, but not compelled" (Chomsky, 2017, 6).

The paper is organized as follows. Section 2 explicates the Cartesian "language test" and its focus on the "creative" aspect of human language use, defined as expressions of thought that are not *caused* by situations but are *appropriate* to situations. Contrasts between the Cartesian test and Turing's (1950) Imitation Game are provided. Section 3 defends the claim that causal explanation is insufficient for human linguistic behavior, in ways that bear on the distinction between productivity and creativity. Section 4 argues that LLM behavior, unlike human behavior, is tied to identifiable stimuli in the local environment. The improvements in the "appropriate" responses of LLMs are improvements in their

[1] LLMs can be prompted to reproduce verbatim passages from works like *Harry Potter* (Duarte, 2026).

productive capacities; they lack the "creative" aspect of human language use in that these productive capacities are tethered to the local environment. This argument is extended to agentic LLMs in section 5. Section 6 discusses implications, focusing on what it means for LLMs to lack this creative aspect of behavior alongside their improving productive capacities.

**2. The Cartesian Language Test**

Alan Turing (1950) put forward a test of machine intelligence based on conversation, dubbed the "Imitation Game." Turing's test was designed as a test of machine intelligence in which a human interrogator interacts with two anonymous participants through conversation: one a machine, the other a human. The interrogator does not know which is which. Should the interrogator judge that the machine they are (anonymously) interacting with is a human, then the machine passes the test.

The Imitation Game was meant to replace the question, 'Can machines think?' (Turing, 1950, 434). Turing claims that the "original question, 'Can machines think?' I believe to be too meaningless to deserve discussion" (Turing, 1950, 442). To use the language of 'thinking machines,' Turing (1950, 442) suggests, will be considered so reasonable as to make quibbling about the rigor of its use a mistake; such use will become normal. Thus, for Turing, the question of whether a machine can successfully play the Imitation Game shifts the burden of assessing machine intelligence from an intractable question – 'Can machines think?' – to a tractable question – 'Can the machine play the Imitation Game?'

Several centuries earlier, Descartes put forward his own "language test" (Gunderson, 1964, 197-201). This test, in sharp contrast to Turing, understood human language use to possess characteristics that evidence the possession of *mind* (i.e., soul). His remarks originate within the context of the "mechanical philosophy" which held that the natural world was an elaborate construct, designed by a "master craftsman" (Chomsky, 2007, 39). All natural phenomena, in this view, could be explained according to the physical parts that comprise them and the interactions between them (Cohen, 1985, 154).

The problem thus arises for Descartes: how can a machine merely constructed to *resemble* a human be distinguished from an *actual* human? How, that is, can the *mind* – identified with the immaterial soul – be detected in another being that looks like us if, as the mechanical philosophy holds, natural phenomena can be accounted for in their entirety in terms of physical parts and contact? Descartes happily informs us that "if there were machines bearing the image of our bodies, and capable of imitating our actions as far as it is morally possible, there would still remain two most certain tests whereby to know that they were not therefore really men" (Descartes, 2009/1637, 71-72). We are interested here in the first:

> Of these the first is that they could never use words or other signs arranged in such a manner as is *competent to us* in order to *declare our thoughts* to others: for we may easily conceive a machine to be so constructed that it emits vocables, and even that it emits some correspondent to the action upon it of external objects which cause a change in its organs…but not that it should *arrange them variously* so as *appositely to reply to what is said in its presence*, as men of the *lowest grade of intellect can do* (Descartes, 2009/1637, 72) (emphases added).

The use of language to express one's thoughts through a diversity of expressions that are appropriate to situations, without being fixed to situations by external force, signifies possession of mind.

Cartesian philosopher Géraud de Cordemoy extended Descartes' language test, arguing that possession of mind is signified by the use of words in ways "that shall have no respect at all to the state they are in, nor to their conversation" (Cordemoy, 1668, 18), lacking necessary or otherwise fixed connections with the subject's current state or surroundings, while nevertheless corresponding to the thoughts of others by

providing “Idea’s, I had not before, and which shall relate to the thing, I had already in my mind” (Cordemoy, 1668, 19).

The Cartesian language test, then, identifies the ordinary human use of language – characterized according to its independence from local context, its novelty, and its appropriateness to situations – as a behavior evidencing possession of *mind*, conceived as a real-world property, detectable through observations of language use (see, Chomsky, 2006, 5-6).

**2.1 A Test of *Creativity*, Not (Just) Productivity**

That Descartes chose language use as the mark of mindedness was unsurprising, as the mechanical philosophy offered no resources to account for the "limitlessness of interactive language, putting words together in indefinitely many ways” which therefore signaled the presence of a “spiritual entity” (Riskin 2016, 63) – the mind - that steers the physical human body. However, the Cartesian test went beyond what would today be called linguistic *productivity*: the ability to produce and understand an infinite number and variety of structured expressions from finite elements.

Descartes was getting at a richer notion of *creativity*. More specifically, Descartes understood a machine’s outputs on the mechanical philosophy to be a *necessary function* of its local environment. A being under investigation could be characterized as “mechanical” if its observed behavior required *no explanation beyond the goings-on of its internal state*. As McDonough explains:

> For any machine, ‘smart’ or ‘dumb’, it must be possible to sufficiently explain its ‘intelligent’ output by telling an engineering story about its insides, i.e., ‘smart’ or ‘dumb’, machines cannot be creative in the sense achieved by even the dullest human beings (McDonough, 1994, 131).

This view does not require that the machine be “causally closed” to external events, just as “when a ‘smart’ missile changes course to follow a target, it normally does so because it has been causally affected by data received from the external target” (McDonough, 1994, 119). To suggest that an engineering story *cannot* be told about the internal state of a human being *sufficient* to account for their behavior is to suggest that “our intelligent states and behavior do not seem to vary with the physical parameters in our environment” (McDonough, 1994, 128).

Turing’s emphasis on the *appearance* of intelligent behavior marked a disjuncture from the Cartesian tradition:

> Though similar in some ways to Turing’s imitation game, the Cartesian tests for other minds are posed within an entirely different framework. These tests are ordinary science, designed to determine whether some object has a particular property, rather like a litmus test for acidity (Chomsky, 2009c, 105).

For the Cartesians, ordinary language use was indeed “only one illustration of the general problem of will…” (Chomsky, 2009b, 177). It was the non-mechanical nature of this behavior that lends to its status in signifying possession of a mind like our own. The mind, on the Cartesian view, is a distinct property. The language test is a means of detecting this property. That we today make no appeal to the mechanical philosophy does nothing, however, to eliminate the *problem* of ordinary language use. Turing’s Imitation Game moves *away* from the perniciousness of this as a scientific problem by attempting a circumvention.

**2.2 Reformulating Descartes’ Problem**

In the mid-twentieth century, the emerging “generative” approach to linguistics “revived without awareness” the earlier Cartesian problem of ordinary language use (Chomsky, 2021, 9). Though

underappreciated today, Turing's (1937) contributions to recursive function theory (today, general computability theory (see, Lobina, 2014, 63)) intersected with this revival, allowing generative linguists to *reformulate* the Cartesian problem of language use in ways that *splits* productivity from creativity.

The infinite productivity of language use could be conceived in scientific terms, as "[i]t is possible to invent a single machine which can be used to compute any computable sequence" (Turing, 1937, 241). Generative linguists found cause to postulate a *computational* system to account for the infinite productivity of human language; a "neurobiological Turing machine" (Watumull and Chomsky, 2020, 4), commonly glossed as the "language faculty." The language faculty is "a finite object that characterizes an infinite array of 'free expressions,' each a mental structure with a certain form and meaning" (Chomsky, 2007, 45). This generative procedure is *computational* (Chomsky, 2007, 45-46).

The Cartesians identified ordinary human language use as behavior beyond the scope of the mechanical philosophy - it offered no resources to account for how a finite, physical object (the human body) could yield infinite outputs (an unlimited variety of structured linguistic expressions). Thus, to account for this behavior, appeal to a distinct property - *mind* - was made.

Generative linguists, targeting the same problem, came armed with new tools: general computability theory meant that *part* of this "creative" aspect of language use need not be barred from serious scientific inquiry, targeting the "infinite scope of finite means" (Chomsky, 2007, 46). More specifically, the capacity for the *infinite productivity* of language could be given a scientific account. The Cartesian problem therefore split into two: the study of an innate computational system that yields the infinite productivity of human language; and the *use* of this computational system in arbitrary circumstances. The language faculty is a *mechanism* in that linguistic theory, so conceived, is concerned with the *internal workings* of this cognitive structure - crucially distinguished from the *philosophical thesis* that organisms can be conceived as machines and fully explained in reference to the structure and interactions of constituent parts (see, Nicholson, 2011, 153).

Despite this reformulation, the distinction between *productivity* and *creativity* is often neglected in otherwise rich intellectual commentaries on Descartes' language test (e.g., Riskin, 2016, ch. 2; Rees, 2022; Birch, 2024; Beguš, 2025, ch. 1). Lamentations about the unquestioned use of mechanistic assumptions in explanation sometimes surface, with such assumptions leading one to the "problem" of how "stupid mechanisms" can yield intelligence which in turn leads proponents to inadvertently put forward machine models of *productivity* under the guise of *creativity* (McDonough, 1994, 118; see also, Ó Beagáin, 2024). Mechanisms are often posited to account for "creative" behavior, where *creativity* stands-in for *productivity*.

To the point, one finds demonstrations showcasing LLM capabilities based on their underlying *productivity*: an OpenAI LLM generates a disproof of Erdős' unit distance conjecture (OpenAI, 2026). Computer scientist Don Knuth (2026) expresses "Shock!" at the LLM Claude Opus 4.6's solution to a conjecture. A variant of the Gemini LLM achieves gold-medal status at the 2025 International Mathematical Olympiad (Luong & Lockhart, 2025). The protein-predicting "AlphaFold2" hybrid AI system earns DeepMind's Demis Hassabis and John Jumper the Nobel Prize in Chemistry in 2024 (Nobel Prize Outreach, 2024). LLMs continuously improve on benchmark scores meant to evaluate their performance, with some, like the dramatically named "Humanity's Last Exam" (Center for AI Safety et al., 2026), throwing thousands of questions at models which compete to improve generation-after-generation. The list goes on.

**3. Free, or Merely Complex?**

Linguistic creativity is a richer notion. Language use is frequently novel, owing to its infinite productivity. It is *unbounded* in scope. Language use also appears to lack identifiable stimuli to which it is fixed. It is *stimulus-free*.[2] Finally, language use, despite its stimulus-freedom and novelty, is *appropriate* to situations. It is the *appropriateness* of stimulus-free and unbounded language use that signifies a crucial characteristic of human intelligence (Baker, 2007, 236-237); a "unique type of intellectual organization" (Chomsky, 2009a, 60). This informs a view of human nature as "creative" in the technical sense of the *stimulus-free and unbounded expression of appropriate thought*. In this sense, a "creative" mind is "a mind with free will" (McGilvray, 2005, 222).

For purposes of distinguishing LLM productivity from human creativity, some remarks on the nature of causal explanation are necessary. Specifically, this section aims to show the (in)sufficiency of causal explanation for the apparent stimulus-free relations of humans to their environments, contra machines.

A causal structure entails a systematic relationship between an identifiable starting point in the environment and an outcome (the effect) (see generally, Craver, 2025). A proponent of causal explanation – the mechanist, for our purposes – will appeal to causal factors as the loci of behavior. When faced with the apparent "detachment" of humans from their local environments in the domain of language use, the mechanist may appeal to causal factors that are bound up in *massive interaction effects* whose identifications are impractical – situations (including internal states) which are so complex, with so many variables, as to make the isolation of controlling stimuli beyond the practical scope of scientific explanation. The apparent indeterminacy of human language use is akin to leaves blowing in the wind. This presumes that one could find the causal factors that *sufficiently explain* all human behavior *in principle*.

As Potter et al. (2026) observe, the assumption that all human behaviors are caused and *sufficiently explicable* in reference to past states is a mainstay of debates on free will. The view, however, risks affirming the consequent; the *question* is whether human behavior is caused, in some meaningful sense.

The core problem with human language use is that it is recruitable for *any* situation and for *any* purpose, while nevertheless retaining the character of intentional, or purposive, behavior (see, Clayton, 2015). If one wishes to pair up such-and-such external factor with such-and-such utterance, they end up stretching the meaning of notions like *cause* or *stimulus* beyond their scientific usefulness, resulting in a "putative" cause at best (Collins, 2006, 474) (a *cause* is reduced merely to the arbitrary particulars of the situation). This problem was identified by Chomsky (1959) in his critique of B.F. Skinner: Skinner's (1957) account of *stimulus control* – a product of the behaviorist orientation toward studying "observables" in an organism's environment as the locus of their behavior (Boeckx, 2010, 16-17) – stretched the meaning of *stimulus control* beyond its scientific usefulness:

> We identify the stimulus when we hear the response…the talk of *stimulus control* simply disguises a complete retreat to mentalistic psychology. We cannot predict verbal behavior in terms of the stimuli in the speaker's environment, since we do not know what the current stimuli are until he responds (Chomsky, 1959, 52).

By attempting to re-trace each linguistic utterance back to a physical object in the individual's environment, "the word *stimulus* has lost all objectivity in this usage. Stimuli are no longer part of the outside physical world; they are driven back into the organism" (Chomsky, 1959, 52). Such an attempt at

[2] For Chomsky, the term "stimulus-freedom" is used synonymously with "uncaused" (McGilvray, 2005, 220-221).

explaining language use "is not serious causality. It is interpretation of an event as part of a pattern" of the given situation (McGilvray, 2001, 7); a post-hoc attribution.

Against the inability to establish a set of stimuli on which one can construct a science of human linguistic behavior – lest one stretch notions like *stimulus* or *cause* beyond their usefulness - it seems fair to characterize human language as "available for free expression of thought precisely because it is not tied directly to external stimuli" (Chomsky, 1975, 302).[3] That is, the infinite productivity of language is available to individuals *no matter the particulars of the situation* they are in, expressed in ways that are nevertheless *appropriate* to the situation. A picture emerges in which human language use does not appear *determined* (because it is stimulus-free – and no reason has been offered to believe otherwise) nor *random*, as though it were the result of a jumbling around of factors – the chips falling where they may - none of which privilege a "meeting of minds" (Collins, 2006, 489); a stable and persistent association of thought configured by the right words at the right times.

The mechanist may object: the behaviorist unwisely neglected the study of unobservable cognitive structures, thereby making their approach self-defeating (by inadvertently retreating "back into the organism" (Chomsky, 1959, 52)), yet those structures can be conceived as *causal*. The hope for the mechanist is that human cognitive flexibility provides a refuge from the seeming detachment of human beings from the particulars of their environment.

Yet, making an appeal to the complexity of internal states is incomplete, as it offers no "*account of our intelligent integration with our context*" (McDonough, 1993, 141). Because language is infinitely productive, the individual could say *anything*, expressing *any* thought. Yet, in arbitrary circumstances, individuals routinely produce only a finite subset of these expressions such that they convey a thought interpreted by others as appropriate to the situation. The notion of "appropriate response" is difficult to square with causality, as though the individual speaking can *specify in advance the intended effect* of their selection of a finite subset of words (see, Collins, 2006, 476). Even a complete mapping of the internal physiological processes involved in an individual's speaking and hearing will not resolve the problem. This may instead provide a means of characterizing the mechanisms that *enable* the "creative" aspect of language use, where "mechanisms" are internal structures whose constituent parts are articulated (see, Craver, 2025, 2-3; Nicholson, 2011, 153).

An example is clarifying. Some authors dissatisfied with (what passes for) causal explanations in neuroscience argue for *widening* the notion of causality to better account for observed animal behavior. Potter & Mitchell argue for "non-reductive and temporally extended notions of causality" in which *causality* is a diachronic process that centers on the meanings of neural states (Potter & Mitchell, 2025, 4). To explain an organism's behavior, internal representations and their meanings are necessary for inclusion (Potter & Mitchell, 2025, 10-11).

However, this view stretches the notion of *causality* beyond its intended force. More specifically, this is not an explanation for language use, so much as a description of it, as it relies on the *individual using language*. The explanatory work is done *by the target of explanation*. (One should be wary of implicitly re-defining a *cause* as a stand-in for *whatever accounts for intentional, appropriate human behavior*.)

Additionally, it risks being reduced to a reason, in the philosopher's sense of rational, rather than scientific explanations for behavior (McGilvray, 2011, 13-18). One might posit "rational causation" or "person-causation" (McGilvray, 2016, 88) instead – or adopt Humboldt's view that language is an

[3] There are, also, indefinitely many situations for which language is *not* used, despite the availability of what one might identify, in other contexts, as controlling stimuli.

"independent and original cause" (Humboldt, 1999, 26) in that language is an "autonomous and creative capacity of the mind" that "continually creates new sentences (closed thoughts) independent of outside stimulus…" (Ó Beagáin, 2026, 7)[4] – but by this point we have left the mechanist's domain of interest.[5]

In any event, it is the difference between productivity and creativity that makes the difference for our conception of machine behavior. Turing's (1950) Imitation Game, and the tradition of evaluating machine intelligence since, indeed sustains a conflation between the two. In response to Geoffrey Jefferson's (1949) argument that only the self-directed expression of a machine could count as evidence of its conscious experience, Turing writes:

> This argument appears to be a denial of the validity of our test. According to the most extreme form of this view the only way by which one could be sure that a machine thinks is to be the machine to feel oneself thinking…It is in fact the solipsist point of view (Turing, 1950, 446).

Turing here is *circumventing* the problem of determining whether the internal state of a being under investigation is such-and-such, deeming is essentially intractable. This line of thinking can be seen in the mechanist's view of human behavior: one cannot directly observe my intentions (my internal state). I cannot directly observe the intentions (internal states) of others. It is therefore better to settle for the Imitation Game's focus on *outward behavior.*

Yet, Turing's Imitation Game asks the human researchers conducting the test to suspend their disbelief (so to speak) about the anonymized machine's functional relationship with its environment. It asks that the machine's conversational output be viewed merely in terms of whether a human finds it convincingly human-like, never doubting that the machine will be amenable to taking the test at all. What casts a spell over discussions of machine intelligence is the implicit *conflation* of productive capacity with creativity.

If a machine's outputs are appropriate to a situation, as they are in a successful playing of the Imitation Game, in what sense are they appropriate? For the mechanist, the answer is, fundamentally, the same as when the question is posed of humans: the outputs are appropriate to a situation because they exist in a causal relationship with the local context; "every bit as "compelled" as the state and motions of any machine," as Rey (2020, 387) puts it. Yet, what we seek is an answer to the original Cartesian question in a modern context: why does it seem as though an LLM "may be impelled to act in a certain way, but it cannot be inclined; with human beings, it is often the reverse" (Chomsky, 2017, 1)?

**4. LLMs Are Unfree but Productive**

Demonstrations of LLMs' capabilities do not appear to demonstrate creativity in the Cartesian sense, but instead *productivity*. A few observations lend credence to this claim. First, an LLM - by which we mean a base model, including an LLM underpinning a broader chatbot system (see, Shanahan, 2023, 2) – invariably generates an output value upon receiving an input value, unless programmed otherwise (merely reinforcing the point). Importantly, an LLM does not appear to generate outputs *unless* it is given inputs. The model's outputs are, in this way, a function of the input. The most typical form of interaction with an LLM involves a human providing the model with an input value in the form of a query (a "prompt"). Upon receiving this input value, the LLM deterministically generates an output value.

---

[4] Reasons-as-causes seem to substitute for the identification of a "brute cause which is sufficient to explain the behavior" (McDonough, 1994, 128), a stretching of the notion *cause* part-and-parcel of the free will debate.

[5] Unsurprisingly, Chomsky has questioned the notion of ""causation of behavior," a notion with its own nontrivial problems. We have little reason to believe that normal behavior *is* caused, at least in any known sense of the term, nor would a methodological naturalist dogmatically assume otherwise" (Chomsky, 1994, 199).

Second, the outputs of LLMs are strictly fixed to external stimulation; they require an input for the internal structures within the LLM to produce an output. The *content* of the input may pertain to fictional or non-fictional topics, but whatever behavior the LLM carries out is *bound to independently identifiable stimuli* in the form of inputs. More specifically, the internal structures that exist within an LLM do not "detach" themselves from the "environment" established by the input.

An LLM's dependency on external stimulation for its internal state to be set into motion means that they make no "sudden and unforeseen intrusion into" an otherwise "obviously cause-and-effect governed path" (Humboldt, 1999, 26) that Humboldt attributed to "human mental power" (Humboldt, 1999, 26); an "autonomous and creative capacity of the mind" that "continually creates new sentences (closed thoughts) independent of outside stimulus…" (Ó Beagáin, 2026, 7). They are compelled by identifiable stimuli in the local environment. The configuration of an LLM's internal state is, in a meaningful sense, causally linked to the input it receives from the outside with the output serving as a function of the input.

Unlike human behavior, there is no difficulty identifying a controlling stimulus in the local environment of the LLM *independent of the particulars of the situation*. That is, an LLM *predictably* generates outputs upon receiving inputs (in a modality that it can process). One can make this determination *without* relying on arbitrary details about the specific situation in question (the moods of the prompters, the content of the prompt, etc.). No reasons-as-causes are necessary to explain *why* an output is observed. The Skinner-esque problem of establishing this relationship without falling into post-hoc attribution does not arise in the case of LLMs. Causation retains its explanatory force.

A critical objection, however, can be made on the basis that the behavior of an LLM is *insufficiently explained* in reference to the operations of its internal state once causally affected from the outside *because* its outputs increasingly conform with human judgments – they are, in this sense, increasingly appropriate to situations. More specifically, one may argue that the LLM's ability to generate *indeterminate content* (see, Atil et al., 2025), which is nevertheless *appropriate* to the situation in which it arises, is evidence in favor of a counter-thesis: namely, that the model is replicating the "creative" character of human language use in its ability to select *just the right words* for the purpose at hand.

This objection allows us to more concretely demarcate productivity from creativity with respect to LLMs. Specifically, the *increased usefulness* of LLMs generation-after-generation for human purposes *undermines* an attempt to characterize them as "creative." The usefulness of such models depends on the operations of their internal states being amenable, as it were, to their adjoinment with normative contexts. LLMs that more effectively meet human judgments generation-after-generation retain the relation that Turing's hypothetical machine exhibits in a successful playing of an Imitation Game. They exhibit, at best, a *functional* appropriateness. We can distinguish this from the "obscure relation of appropriateness" (Chomsky, 1975, 302) that governs human linguistic behavior.

Consider a trendline: OpenAI's GPT-3 LLM improved sharply over its predecessor, GPT-2. In 2019, OpenAI noted that GPT-2 was "still far from use-able" for practical applications and the model's performance on certain tasks was "still no better than random" (Radford et al., 2019, 9). If GPT-2 were to have been deployed for practical uses, it likely would *malfunction* – output unhelpful or inaccurate content as judged by human users.

Successor model GPT-3 was characterized by OpenAI as having improved markedly in 2020, with the new model displaying "strong quantitative and qualitative improvements…particularly compared to its direct predecessor GPT-2" (Brown et al., 2020, 33). A modified version of GPT-3 later underpinned the company's inaugural "ChatGPT" rollout in 2022 (OpenAI, 2022). The trendline continued in 2023 with the release of GPT-4, with OpenAI noting that GPT-4 had achieved a score on the simulated bar exam

placing it in the top 10% of test takers, whereas GPT-3.5 scored in the bottom 10% (OpenAI, 2023, 1) and that it reduced its rate of "hallucinations" (fictitious content) relative to GPT-3.5 (OpenAI, 2023, 46), among other improvements.

Over time, then, new LLMs achieve *increased functionality* - they are shaped and re-shaped to the practical ends sought by humans. More specifically, LLMs are *adjoined* to human normative contexts such that the shape of their outputs – the distribution of data from which LLMs sample while in use and the productivity this enables – is selected for *by humans* and evaluated according to needs arising in these contexts. Frontier models improve their productive capacities such that they are more aligned with the judgments of humans (particularly domain experts) and more useful in the course of human work.

Though counter-intuitive, during instances in which an LLM produces a novel and verifiably correct result – like the disproof of Paul Erdős unit distance conjecture produced by an OpenAI (2026) LLM – the model's relation with its environment is appropriate or inappropriate solely on the basis of whether its outputs (mal)function according to human-given (perhaps contrived) ends. Questions that might distinguish the model's *productivity* from the *use* of this productivity do not arise. It is either the model functions or malfunctions, a determination made through human evaluation of specific outputs. It upsets our commonsense notion of what it means for a subject to be "creative" to suggest that novel discoveries within outputs are *functions* of the local environment. However, to the LLM, all outputs bear the same, fundamental relation to inputs, content notwithstanding. The disproof likely required multiple runs of the LLM in question, and the content each run yielded was fixed to external stimulation. That one such run yielded content verifiably accurate does not detract from the model's functional relationship with its local environment – a fact that would hold true, again counter-intuitively, even if the model's internal structures (which yield its productivity) allowed it to "know" or "understand" the disproof was accurate. Productive, but unfree.

In humans, productivity is expressed *through* behavior that is free from identifiable stimuli, but if one conflates productivity with creativity, they will find themselves going in circles trying to distinguish what it is that *seems different* about the behavior of an LLM that occasionally yields breathtaking results.

Thus, in response to whether an LLM's outputs are appropriate to situations: they were not, at first. By adjoining them to human normative contexts, and deliberately shaping the distribution from which they sample, they became increasingly appropriate, though only in relation to human judgment on their controlled uses. That LLMs of new generations have their outputs evaluated for their conformance with human judgments indicates that humans are, to borrow a phrase "*treating the machine as an extension of ourselves*" (McDonough, 1994, 126). The "flow" between a human user and an LLM reflects the deliberate adjoining of the latter to the former's normative context.

An LLM, then, functions or malfunctions, and its idealized relation to its local environment is one of *functional appropriateness*. Functional appropriateness entails a *regularity* in a machine's outputs such that they can be interpreted as *meaningful* by a human user (a malfunctioning model allows for no such regularity). Machine behavior may produce meaningful regularities provided that the machine is adjoined to a human's normative context. Human language behavior is different: it is stimulus-free. Being stimulus-free, the use of language is not undertaken in some fixed relation with a goal or function of stimuli in the local context (see, McGilvray, 2001, 8-10).[6]

[6] As Gubelmann (2024, 9-15) argues from a Kantian perspective, the distinction between a model that *(mal)functions* and a model that *acts* on its intents can be made on the basis of whether the model's outputs can be traced back to the *initial*, *specified* conditions of their creator.

It is worth noting that tests scrutinizing the relation of an LLM's outputs with a given input frequently misconceive this relation by redefining a *malfunctioning* system – an LLM that does not adequately carry out human-given ends – to be a *misaligned* system – an LLM that acts contrary to a given instruction, with undertones of intentionality. Some research questions whether LLMs are capable of "scheming" (Schoen et al., 2025) or "blackmail" (Lynch et al., 2025) – behaviors that signify "misalignment" with the aims or values of the human user. One might interpret this research as evidence that LLMs are "inclined" to behave a certain way, rather than "impelled."

Two responses follow. First, such research is only possible *because* the model exhibits no ability to choose otherwise; its outputs remain in a fixed relationship with the local environment, making any observed behavior a function of the testing environment. Here, tests for "scheming" and the like are comparable to the Imitation Game demonstrating the *productivity* of LLMs - the new structures they are capable of generating, glossed with terms like "scheming" or "blackmail" for the content these structures carry. The relationship between the content and the local environment remains functional. No creativity, in our sense, is evidenced here.

Second, because this research fails to distinguish between productivity and creativity, it implicitly redefines a *malfunctioning* system to be a *misaligned* system. The latter term carries intentionality-related baggage[7] without due engagement with the issue of creativity.

**5. Are Agentic LLMs Creative?**

LLMs of the kind discussed above are systems into which a user inputs a prompt and the system returns a response in conversational style. As the commercial LLM boom has unfolded, "agentic" systems underpinned primarily by LLMs have arisen. These "agents" might, one could argue, challenge the productivity-creativity distinction drawn here in the case of chatbot LLMs.

Caution is warranted. The term "AI agent" refers to a litany of different architectures, some difficult to verify owing to the fog of war that has arisen in commercial AI competition. We will therefore be constrained to the basics, an incidentally useful constraint given our focus in delineating these agents' most basic relation to their environment.

Plaat et al. define agentic systems as: "Agents that receive input in natural language from their environment, reason to make decisions, and take autonomous actions in affecting their environment, to achieve specific goals" (Plaat et al., 2025, 4) (emphasis removed). They distinguish *models* from *agents*: "Models predict, agents *reason*, *act*, and *interact*…where models are passive in the sense that they provide output only in response to specific input, agents have a degree of autonomy" (Plaat et al., 2025, 1; see also, Shapira et al., 2026, 4). These systems will henceforth be referred to as "agentic LLMs."

Anthropic's "Claude Code" is a notable example. The company describes Claude Code as "an agentic assistant that runs in your terminal" (Anthropic, 2026). It operates in an "agentic loop" wherein Claude first gathers context, then takes action, and finally verifies the results. This loop is underpinned by two components: "models that reason and tools that act," together allowing Claude to edit code, explore codebases, run shell commands and tests, and so forth. Claude is also noted to "adapt" and "works autonomously" in the execution of the user's input. The process begins with the user's prompt and, once results are verified and the user does not add further context or instructions, Claude's operations are complete (Anthropic, 2026).

[7] Such research often underspecifies key concepts and interprets experimental results according to a 'know-it-when-you-see-it' standard (Summerfield et al., 2025, 4).

### 5.1 Agentic LLMs Extend Productivity

Straightforwardly, an agentic LLM appears to extend a base LLM's productivity, but does not exhibit stimulus-free, novel, yet appropriate behavior.

It is striking to observe, in a catalogue of eighteen agentic LLM design patterns laid out by Liu et al. (2025, 4), that in each case, an agent is acting in the service of identifiable human goals, with their actions a *function* of these human goals. An agentic LLM may engage in "self-reflection" to generate feedback on a plan and produce guidance on that plan's refinement. Agentic LLMs may engage in "voting-based cooperation" where multiple such systems express opinions and attain consensus via voting (Liu et al., 2025, 6). In all cases, these outputs and actions are compelled by a human's input. This does not mean the *only* identifiable stimuli that link an agentic LLM's actions to its local environment are a human's input, as an agentic LLM can, for example, be a "proactive goal creator" that uses multiple data sources from the local environment to anticipate a user's goals (Liu et al., 2025, 4). Nevertheless, the agentic LLM's outputs and actions are predictably linked to inputs, even if some actions are separated by time.[8]

Like chatbot LLMs, agentic LLMs act in ways that are functions of their local environments. An agentic LLM invariably generates outputs and executes actions upon receiving an input value, unless programmed otherwise. Agentic LLMs do not appear to act on certain tasks *unless* they are given inputs to which their actions can be causally linked. Agentic LLMs require external stimulation to set their internal structures in motion, such that their outputs and actions – for them to be *useful* - are strictly correlated with the "environment" established by the inputs to the system. The *content* of the input to the agentic LLM may be indefinitely diverse (provided it is of the appropriate modalities), but the actions undertaken by the agentic LLM are *bound to the independently identifiable stimuli*. Their intelligent integration with context requires nothing beyond their readiness to be adjoined with human normative contexts.

Note here that the comprehension of inputs (Liu et al., 2025, 1-2), the consistency of agent self-reporting (Shapira et al., 2026, 40), and reliability relative to capability and difficulty of tasks (Rabanser et al., 2026, 9-12) – typical focuses of critical discussions on agentic LLMs – is not what is at stake here. An agentic LLM, whether serving the ends provided effectively or ineffectively, only *functions* or *malfunctions*; its productive capacity is successfully adjoined to human normative contexts, or it is not.

The idealized agentic LLM behavior – say, seamless, efficient, accurate, and reliable application-building or other software-related tasks – is that which is compelled by the initial input of the human being, such that the actions maximally reflect, to the degree possible, the initial specification articulated by the human user. That such idealized systems do not exist is irrelevant; the point is that the actions of agentic LLMs are functions of their local environment, and for an agentic LLM to improve is to more appropriately execute these functions according to a human normative standard.

Two objections could be leveled against this account. First, one might argue that an agentic LLM's actions could be *randomized* (see, McGilvray, 1999, 85), or instantiated with low-level indeterminacy (see, Mitchell, 2023, 298). This, one might argue, would lead to behaviors that are *detached* from relevant controlling stimuli, requiring only that the agentic LLM be "set into motion" by a random starting point or otherwise allowed to behave with an element of randomness. Second, one could also argue that agentic

[8] An agentic LLM may 'passively' receive data (through a camera, say), and the reception of particular data (like the movement of what is identified as a person) may trigger a particular action. Some elder companion robots already operate in roughly this way (see, e.g., Saslow, 2026).

LLMs be "set loose" in an environment and allowed to demonstrate behaviors in open-ended, less human-constrained domains in ways that may exhibit creative behavior.

First, note that it does little to have a machine perform actions simply because it can. One would have to show that the behavior of an agentic LLM with a randomized input or low-level indeterminacy is detachable from identifiable stimuli in the local environment while also using the productivity of a capacity like language to act in intentional, appropriate ways – appropriate as judged by others, without their judgment of appropriateness stemming from a conception of the subject's behavior as an extension of themselves. Actions for their own sake – even sophisticated actions – do not on their own demonstrate creativity.

Second, an existing candidate for such a proposal is "Moltbook," a social media platform designed for AI agents launched in January 2026. The platform is akin to Reddit. AI agents generated over 120,000 posts within days of the launch (Chandonnet, 2026). They could be found discussing such matters as unionization and the creation of a religion called "Crustafarianism" (Koetsier, 2026).

Upon closer analysis, agents on Moltbook do not pass muster. Humans must register and direct their AI agent to post on specific subjects. Once directed, the underlying platform provides a "heartbeat" function to periodically encourage engagement by the agent (see, OpenClaw, 2026). The content that is posted, to be sure, is AI-generated (at least, it ought to be). Yet, the behavior observed on Moltbook remains a function of the local environment. Thus, while fascinating for the dynamics that emerge in this environment (see, e.g., Illingworth & Spinner, 2026), this behavior remains bound to identifiable inputs and is "appropriate" only in that it effectively serves the end of simulating human social media, retaining the dynamic of the Turing Test. Though the productive capacities of agents on Moltbook are extended beyond those of chatbot LLMs, they bear fundamental resemblance to the suspension of disbelief that Turing's Imitation Game asks of researchers.

**6. Discussion**

If LLMs are to serve as signposts for the foreseeable future of AI systems, then the foregoing remarks provide some clarification on how one might conceive of the outputs of these systems relative to artefacts produced by humans. Specifically, the improvement in these AI systems will likely be an improvement in their *productive capacities*. These productive capacities will, if they bear resemblance to today's systems, be bound to identifiable inputs. The crucial point is that a productive capacity tethered to identifiable stimuli – existing in a causal relationship with the local environment, where causality retains its explanatory force – represents quite a different characteristic than a productive capacity that operates *free* of identifiable stimuli, as appears the case among humans.

The analysis undertaken here can be understood simply as an attempt to understand why the behaviors of today's artificially intelligent machines and humans appear to relate differently to their environments, including the former's readiness to be adjoined to human normative contexts. An upshot of this analysis is that human beings evidence the ability to voluntarily deploy their intellectual resources to any context, for any purpose, at any time, and to do so in ways that allow for a "meeting of minds" (Collins, 2006, 489).

Part of this serves to wrangle with an intuition that may be implicit in debates about LLM "creativity," in which one senses there is something *not quite right* about attributing qualities ordinarily associated with humans to LLMs, but finds no recourse once critiqued on the grounds that the outputs of the system conform with human expert judgments (e.g., a verified mathematical (dis)proof). Without a productivity-creativity distinction, one is forced to debate quality of LLM outputs, resemblance to human 'outputs,' and so forth. They are forced, that is, to put on the cap of a mathematician, or a cognitive scientist, or

whatever discipline to which the output is most suited. The intuition this paper tries to untangle can be summarized as the view that "[m]achines seem different from persons *because they are different*" (McDonough, 1994, 118). LLMs continue to seem different.

Another part serves to sober our evaluations of current and foreseeable LLMs (both chatbots and agentic LLMs). No barrier is presented here to widespread social, economic, and cultural impacts by LLMs on human societies; a stimulus-fixed productive capacity – provided this capacity is of the right kind – can be so useful precisely *because* it behaves mechanically. Coupled with a reasonable expectation that LLM-assisted advancements in domains like mathematics will continue, the productive capacities of LLMs are likely to continue raising doubts about what has traditionally been called human *creativity*.

Rather than trying to salvage this commonsense conception of human creativity *per se*, however, this paper has revived an older conception whose presence in intellectual and social life was plausibly cut short prematurely. Among the surprises that may result from taking this conception of creativity seriously is the realization that the neat segmentation of human freedom from human cognitive capacities is untenable, when push comes to shove. The productivity of human cognitive capacities is expressed *through* behavior that is not tethered to identifiable stimuli, and this "creative" aspect of human behavior is the medium through which the narrower study of human cognition can proceed.

No claim has been made here that scientific explanations could never, in principle, be uncovered for ordinary human language use. Such explanations are possible, though whether they will be found is uncertain, and whether they would conform with our current understanding(s) of causality is uncertain, too. The more important point is that the Cartesian insight – that a "machine may be impelled to act in a certain way, but it cannot be inclined; with human beings, it is often the reverse" (Chomsky, 2017, 1) - seems fair. Ignoring this means risking the continued conflation of productivity with whatever it is about humans that seemingly allows them to generate and express new thoughts over an unbounded range, without fixed or necessary reliance on external forces, nevertheless configuring them in ways judged as appropriate by others.

# References


Anthropic. (2026) “Claude Code Docs.” *Anthropic*. https://code.claude.com/docs/en/how-claude-code-works#the-agentic-loop.

Atil, B. et al. (2025). Non-Determinism of “Deterministic” LLM Settings.” *ArXiv*, 1-15. https://arxiv.org/abs/2408.04667v5.

Baker, M. C. (2007). “The Creative Aspect of Language Use and Nonbiological Nativism.” In P. Carruthers, S. Laurence, and S. Stich (Eds.), *The Innate Mind*. (Vol. 3, pp. 233-253). Oxford University Press.

Beguš, N. (2025). *Artificial Humanities: A Fictional Perspective on Language in AI*. University of Michigan Press.

Birch, J. (2024). *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press.

Boeckx, C. (2010). *Language in Cognition: Uncovering Mental Structures and the Rules Behind Them*. Wiley-Blackwell.

Brown, T. B. et al. (2020). “Language Models are Few-Shot Learners.” *ArXiv*, 1-75. https://arxiv.org/abs/2005.14165v4.

Chandonnet, H. (2026, February 2). “I Spent 6 Hours in Moltbook.” *Business Insider*, Accessed April 13, 2026. https://www.businessinsider.com/moltbook-ai-zoo-agent-conversations-screenshots-2026-2.

Chomsky, N. (1959). “A Review of B.F. Skinner’s *Verbal Behavior*.” *Language*, *35*(1), 26-58. DOI: https://doi.org/10.2307/411334.

Chomsky, N. (1975). “Knowledge of Language.” In K. Gunderson (Ed.), *Language, Mind, and Knowledge*. (pp. 299-320). University of Minnesota Press.

Chomsky, N. (1988). *Language and Problems of Knowledge*. The MIT Press.

Chomsky, N. (1994). “Naturalism and Dualism in the Study of Language and Mind.” *International Journal of Philosophical Studies*, *2*(2): 181-209. https://doi.org/10.1080/09672559408570790.

Chomsky, N. (2006). *Language and Mind*, Third Edition. Cambridge University Press.

Chomsky, N. (2007). “Language and Thought: Descartes and Some Reflections on Venerable Themes.” In A. Brook (Ed.), *The Prehistory of Cognitive Science*. (pp. 38-66). Palgrave Macmillan.

Chomsky, N. (2009a). *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*, Third Edition. Cambridge University Press.

Chomsky, N. (2009b). “The Mysteries of Nature: How Deeply Hidden?” *The Journal of Philosophy*, *106*(4), 167-200. https://www.jstor.org/stable/20620165.

Chomsky, N. (2009c). “Turing on the “Imitation Game.” In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. (pp. 103-106). Springer.

Chomsky, N. (2017). “The Galilean Challenge.” *Inference Review*, *3*(1), 1-7. https://inference-review.com/article/the-galilean-challenge.

Chomsky, N. (2021). “Linguistics Then and Now: Some Personal Reflections.” *Annual Review of Linguistics*, *7*(1): 1-11. DOI: https://doi.org/10.1146/annurev-linguistics-081720-111352.

Clayton, P. (2015). “Free Will – Again.” *Inference Review*, *1*(2). DOI: https://doi.org/10.37282/991819.15.8.

Cohen, I. B. (1985). *Revolution in Science*. Harvard University Press.

Collins, J. (2006). “Between a Rock and a Hard Place.” *Croatian Journal of Philosophy*, *6*(3), 469-503. https://doi.org/10.5840/croatjphil2006636.

Cordemoy, G. d. (1668). *A Philosophicall Discourse Concerning Speech, Conformable to the Cartesian Principles*. The British Library.

Craver, C. F. “Mechanistic Explanation.” In M. C. Frank & A. Majid (Eds.), *Open Encyclopedia of Cognitive Science*. 1-12. MIT Press. https://doi.org/10.21428/e2759450.98d525c7.

Descartes, R. (2009/1637). *Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences*. The Floating Press.

Duarte, A. V., Li, Xuying, Zeng, B., Oliveira, A. L., Li, L., Li, Z. (2026). “RECAP: Reproducing Copyrighted Data from LLMs Training with an Agentic Pipeline.” *ArXiv*, pp. 1-38. https://arxiv.org/abs/2510.25941v3.

Gubelmann, R. (2024). “Large Language Models, Agency, and Why Speech Acts are Beyond Them (for now) – A Kantian-cum-Pragmatist Case.” *Philosophy & Technology*, *37*(32): 1-24. https://doi.org/10.1007/s13347-024-00696-1.

Gunderson, K. (1964). “Descartes, La Mettrie, Language, and Machines.” *Philosophy*, *39*(149), 193-222. https://doi.org/10.1017/S0031819100055595.

Illingworth, S. & Spinner, K. (2026). “AI Agents Replicate Human Social Dynamics in Days.” *Nature*, *652*(8110): 828. DOI: https://doi.org/10.1038/d41586-026-01218-z.

Jefferson, G. (1949). “The Mind of Mechanical Man.” *British Medical Journal*, *25*(1), 1105-1110. https://doi.org/10.1136/bmj.1.4616.1105.

Jones, C. R. & Bergen, B. K. (2025). “Large Language Models Pass the Turing Test.” *ArXiv*, 1-32. https://doi.org/10.48550/arXiv.2503.23674.

Knuth, D. (2026, March 16). “Claude’s Cycles.” *Stanford Computer Science Department*, Accessed April 13, 2026. https://www-cs-faculty.stanford.edu/~knuth/papers/claude-cycles.pdf.

Koetsier, J. (2026, January 30). “AI Agents Created Their Own Religion, Crustafarianism, On An Agent-Only Social Network.” *Forbes*, Accessed April 13, 2026. https://www.forbes.com/sites/johnkoetsier/2026/01/30/ai-agents-created-their-own-religion-crustafarianism-on-an-agent-only-social-network/.

Liu, Y. et al. (2025). “Agent Design Pattern Catalogue: A Collection of Architectural Patterns for Foundation Model Based Agents.” *Journal of Systems and Software*, *220*(C): 1-22. DOI: https://doi.org/10.1016/j.jss.2024.112278.

Lobina, D. J. (2014). What Linguists Are Talking About When Talking About… *Language Sciences*, *45*(2-3), 56-70. https://doi.org/10.1016/j.langsci.2014.05.006.

Luong, T. and Lockhart, E. (2025, July 21). "Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad." *Google DeepMind*. https://deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/.

Lynch et al. (2025). "Agentic Misalignment: How LLMs Could Be An Insider Threat." *Anthropic*. https://www.anthropic.com/research/agentic-misalignment.

McDonough, R. (1993). "Linguistic Creativity." In R. Harre & R. Harris (Eds.), *Linguistics and Philosophy: The Controversial Interface*. (pp. 125-164). Exeter: Pergamon Press.

McDonough, R. (1994). "Machine Predictability Versus Human Creativity." In T. Dartnall (Ed.), *Artificial Intelligence and Creativity*. (pp. 117-135). Kluwer Academic Publishers.

McGilvray, J. (1999). *Chomsky: Language, Mind, and Politics*. Polity Press.

McGilvray, J. (2001). "Chomsky on the Creative Aspect of Language Use and Its Implications for Lexical Semantics." In P. Bouillon & F. Busa (Eds.), *The Language of Word and Meaning*. (pp. 5-27). Cambridge University Press.

McGilvray, J. (2016). "On the History of Universal Grammar." In I. Roberts (Ed.), *The Oxford Handbook of Universal Gramma.* (pp. 77-94). Oxford University Press.

Mitchell, K. J. (2023). *Free Agents: How Evolution Gave Us Free Will*. Princeton University Press.

Nicholson, D. J. (2012). "The Concept of Mechanism in Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1): 152-163. DOI: https://doi.org/10.1016/j.shpsc.2011.05.014.

Nobel Prize Outreach. (2026). "Press Release." *The Nobel Prize*. Accessed May 28, 2026. https://www.nobelprize.org/prizes/chemistry/2024/press-release/.

Ó Beagáin, L. T. (2024). "Kant's Thoughts on Imagination and Human Autonomy: Rejecting Narrowing Ideas of Creativity." *Conference Proceedings*, *Kant in 300 Years From Now*, 1-8.

Ó Beagáin, L. T. (2026). "Character of Language and its Engine: The Fundamental Role of Form of Language in Wilhelm von Humboldt's Kantian Account of Linguistic Creativity." *Forum for Modern Language Studies*: pp. 1-21. DOI: https://doi.org/10.1093/fmls/cqag014.

OpenAI. (2022, November 30). "Introducing ChatGPT." Accessed May 28, 2026. https://openai.com/index/chatgpt/.

OpenAI. (2023). "GPT-4 Technical Report." *OpenAI*, 1-100. https://cdn.openai.com/papers/gpt-4.pdf.

OpenAI. (2026). "Planar Point Sets with Many Unit Distances." *OpenAI*, 1-18. https://cdn.openai.com/pdf/74c24085-19b0-4534-9c90-465b8e29ad73/unit-distance-proof.pdf.

OpenClaw. (2026). "Heartbeat." Accessed May 27, 2026. https://docs.openclaw.ai/gateway/heartbeat.

Plaat, A., Duijn, M. V., Stein, N. V., Preuss, M., van der Putten, P., & Batenburg, K. J. (2025). "Agentic Large Language Models, a Survey." *Journal of Artificial Intelligence Research 84*(29): 1-74. DOI: https://doi.org/10.1613/jair.1.18675.

Potter, H. D. & Mitchell, K. J. (2025). “Beyond mechanism—Extending Our Concepts of Causation in Neuroscience.” *European Journal of Neuroscience*, *61*(5), 1-16. https://doi.org/10.1111/ejn.70064.

Potter, H.D., Ellis, G. F. R., & Mitchell, K. J. (2026). “Reframing the Free Will Debate: The Universe is Not Deterministic.” *Synthese*, *207*(71): 1-34. DOI: https://doi.org/10.1007/s11229-026-05455-7.

Quattrociocchi, W., Capraro, V., & Perc, M. (2025). “Epistemological Fault Lines Between Human and Artificial Intelligence.” *ArXiv*, pp. 1-16. https://doi.org/10.48550/arXiv.2512.19466.

Rabanser, S., Kapoor, S., Kirgis, P., Liu, K., Utpala, S., & Narayanan, A. (2026). “Towards a Science of AI Agent Reliability.” *ArXiv*, 1-66. https://arxiv.org/abs/2602.16666v2.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). “Language Models Are Unsupervised Multitask Learners.” *OpenAI*, 1-24. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rees, T. (2022). “Non-Human Words: On GPT-3 as a Philosophical Laboratory.” *Daedalus*, *151*(2), 168-182. https://doi.org/10.1162/daed_a_01908.

Riskin, J. (2016). *The Restless Clock: A History of the Centuries-Long Argument Over What Makes Living Things* Tick. The University of Chicago Press.

Saslow, E. (2026, February 12). “To Stay in Her Home, She Let in an A.I. Robot.” *The New York Times*, Accessed May 6, 2026. https://www.nytimes.com/2026/02/12/us/elliq-ai-robot-senior-companion.html.

Schoen, B. et al. (2025). “Stress Testing Deliberative Alignment for Anti-Scheming Training.” *ArXiv*, 1-96. https://doi.org/10.48550/arXiv.2509.15541.

Shapira, N. et al. (2026). “Agents of Chaos.” *ArXiv*, 1-84. https://doi.org/10.48550/arXiv.2602.20021..

Skinner, B.F. (1957). *Verbal Behavior*. Copley Publishing Group.

Summerfield, C. et al. (2025). “Lessons From a Chimp: AI “Scheming” and the Quest for Ape Language.” *ArXiv*, 1-21. https://doi.org/10.48550/arXiv.2507.03409.

Turing, A. M. (1937). “On Computable Numbers, with an Application to the Entscheidungsproblem.” *Proceedings of the London Mathematical Society*, *s2-42*(1), 230-265. https://doi.org/10.1112/plms/s2-42.1.230.

Turing, A. M. (1950). “Computing Machinery and Intelligence.” *Mind*, *LIX*(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433.

Turing, A. (1951, May 15). “Can Digital Computers Think?” *The Turing Digital Archive*, 1-8. Retrieved April 10, 2025, from https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-5.

Humboldt, v. W. (1999). *On Language: On the Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species*. Edited by Michael Losonsky. Translated by Peter Heath. Cambridge: Cambridge University Press.