

Article

Consummatum Est: On the Faustian Laboratory, the Hubris of AI's Architects, and the Humanity They Forgot to Ask

Alison L. Kahn

School of Social Sciences and Humanities, Loughborough University, Loughborough LE11 3TU, UK;
a.kahn@lboro.ac.uk

Abstract

The development of artificial intelligence has proceeded within a disciplinary culture whose positivist epistemological foundations generate a constitutive blind spot: the systematic exclusion of tacit, embodied, relational and contextually situated knowledge from what counts as knowledge at all. This article argues that this exclusion is not a technical limitation, but a structural condition of AI systems as currently built, with serious consequences for individuals, institutions and the social order. The method is interdisciplinary and critically synthetic, integrating the anthropology of science and technology, the philosophy of language and the literary-philosophical tradition of techno-critique. Three original contributions are advanced. First, the Faustian framework is reformulated as a collective rather than individual pact: the AI contract implicates a civilisation. Second, the prevalent reductionist critique is refined: the error is not quantification of incommensurable goods but the enforcement of total orderings upon value landscapes that admit only partial orderings. Third, the stochastic parrot objection is reconciled with the Faustian analysis through the concept of institutional amplification: AI's danger resides not in its intelligence but in the authority conferred upon its incomprehension. The article concludes that Mephistopheles, not Faustus, is the more precise figure for AI itself, and asks whether the defunding of humanities disciplines, those best placed to navigate the moral challenges AI presents, constitutes the most consequential characteristic deletion of all.

Keywords: artificial intelligence; technolatr; epistemology; positivism; Faustian pact; partial ordering; embodied knowledge

'Why, this is hell, nor am I out of it.'

Mephistopheles, Doctor Faustus, Marlowe c.1592, Act 1, Scene 3

'Consummatum est.'

Faustus, Doctor Faustus, Marlowe c.1592, Act 2, Scene 1



Academic Editor: Marcin J. Schroeder

Received: 14 April 2026

Revised: 2 July 2026

Accepted: 4 July 2026

Published: 13 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

The critique of artificial intelligence as a form of technolatr, the elevation of technical capacity to a position of unquestioned cultural authority, has a substantial scholarly history. Langdon Winner's *Autonomous Technology* [1] (p. 15) established the foundational framework: technical systems acquire momentum that outstrips the intentions of their designers, structurally deferring governance until it becomes urgent. Albert Borgmann's 'device paradigm' notes that modern technology makes goods and services 'instantaneous, ubiquitous, safe, and easy' [2] (p. 41) identified the concealment of machinery behind simplified interfaces as characteristic of modern technology's relationship to its users. Neil Postman's *Technopoly* [3] (p. 4) argued that the dominance of technological thinking constitutes a

form of cultural totalitarianism in which no alternative standards of judgement survive. The Frankenstein archetype, identified by Winner as a persistent figure in this tradition, has structured accounts of artificial intelligence from early robotics debates through to contemporary discussions of artificial general intelligence, its logic, the creator destroyed by the created, proving durable because it captures the asymmetry between the time required to build a system and the time available to understand its consequences.

The epistemological dimension of this critique is the question not merely of what AI systems do but of what kind of knowledge they represent and what kind they exclude, which has been developed most rigorously within the sociology and anthropology of science and technology. Michael Polanyi's foundational account of tacit knowing [4], "we know more than we can tell," established the theoretical problem: much of what experts know cannot be articulated in propositional form [4]. Harry Collins and Robert Evans's work on 'expertise and experience' [5] extended this into a framework for distinguishing forms of knowledge that AI systems cannot capture. Lucy Suchman's *Plans and Situated Actions* [6] (p. 27) demonstrated empirically that the computational models of human action underlying early AI were systematically unable to account for the improvisational, context-dependent character of actual human work. Diana Forsythe's ethnographic fieldwork inside AI laboratories from the late 1980s until her death in 1997, collected in *Studying Those Who Study Us* [7] (p. 33), provided the most direct account of what she termed the 'characteristic deletions' of knowledge engineering: the systematic exclusion of tacit, relational and embodied expertise by researchers who were, she argued, typically positivist in their epistemological assumptions.

Her work remains substantially unengaged by the deep learning community, despite its direct relevance to the foundational assumptions of large language model development. The philosophy of language has approached related questions through embodied and enactive accounts of cognition. Maurice Merleau-Ponty's *Phenomenology of Perception* (original in French 1945) [8] established the body as the primary site of world-engagement, stating, "The world is not what I think, but what I live" (2012 p. lxxx, Landes translation of original text) [9] against the Cartesian separation of thinking substance from extended substance that underwrites the computational model of mind. Francisco Varela, Evan Thompson and Eleanor Rosch's *The Embodied Mind*: 1991 [10] synthesised phenomenological and cognitive science perspectives to argue that cognition is inseparable from embodied action. They argue:

"By using the term embodied we mean to highlight two points: first, that cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second, that these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological, and cultural context." [10] (p. 173)

Julia Kristeva's distinction between the symbolic and the semiotic offers a specific account of the bodily, pre-linguistic ground of language, the chora of drives and rhythms from which meaningful speech emerges, that bears directly on what large language models cannot possess and what a differently constituted future system might. She defines it as "an essentially mobile and extremely provisional articulation" [11] (p. 26).

These philosophical resources have rarely been brought into direct dialogue with the empirical STS literature on AI knowledge engineering, and one contribution of this article is to perform that synthesis. The literary-philosophical tradition of AI critique has characteristically employed Marlowe's *Doctor Faustus* [written in 1592, published in 1604] and Shakespeare's *King Lear* [written in 1604, first published in 1608] as its primary figures.

Building on earlier critiques of technological determinism, Winner, 1977 [1] and N. Katherine Hayles demonstrated how late twentieth-century cybernetics came to privilege

“informational pattern over material instantiation,” treating embodiment as “an accident of history rather than an inevitability of life” (2: 1999 Hayles) [12]. Hayles’ analysis traces the historical conditions under which information came to be understood as separable from its material substrate. Kate Crawford’s *Atlas of AI* (2021) [13], situated AI within the political economy of extraction and labour, contributes to this tradition. She states, “Rather than seeing AI as an abstract technical system, we need to understand it as an extractive industry.” [13] (p. 15). The present article positions itself explicitly within it while proposing three refinements that change the diagnosis. These are not cosmetic adjustments: each refinement alters what the diagnosis implies about remedy and governance, and each addresses a specific tension in the existing literature that has remained unresolved.

This paper establishes the Faustian framework and its required collective reformulation, arguing that the AI pact implicates a civilisation rather than merely its named architects. It then traces the positivist tradition from Descartes through Michael Strevens’ iron rule of explanation as the epistemological condition that made AI laboratories structurally resistant to humanistic critique and reformulates the *King Lear* critique at the level of ordering rather than measurement. Drawing on Kristeva’s symbolic/semiotic distinction to specify what is missing from current large language models, while acknowledging that the critique is falsifiable and that differently embodied future systems might satisfy it in different degrees, it integrates the stochastic parrot objection through the concept of institutional amplification, resolving its apparent tension with the Faustian analysis. Mephistopheles is proposed as the more precise literary figure for AI itself, raising the question of demonic versus angelic AI, both within the framework of the pre-Copernican hierarchy of intelligences as a historical precedent for living alongside superior minds, and finally drawing together the argument’s strands, which state implications for governance and for the institutional future of the humanities.

2. Signed in Blood

What the Architects of AI Chose, and What They Chose to Ignore

As Will Douglas Heaven (2023) [14] recounts following Hinton’s departure from Google, Hinton had once explained his persistence in AI research despite his anxieties with the remark: “I could give you the usual arguments. But the truth is that the prospect of discovery is too sweet.” [15] (p. 78). This sentence belongs not to the literature of technology but to the much older literature of hubris, and that older literature has already been brought to bear on artificial intelligence at length and with considerable force, a fact which any honest engagement with the field’s critics must acknowledge before proceeding to whatever it proposes to add. The Faustian figure has served as a touchstone for technology critique since Winner (1977) [1], which drew on the myth to characterise the self-amplifying momentum of technical systems that outrun the intentions of their designers. Mary Shelley’s *Frankenstein* (1818), itself a Faustian narrative, has provided an alternative but structurally identical frame for accounts ranging from early robotics debates to contemporary discussions of AGI. Shakespeare’s *King Lear* has been deployed, most prominently in the philosophy of value and in critiques of economic reductionism, to illustrate the category error of submitting incommensurable goods to a common metric. The present article does not contest the validity of these prior deployments. It argues, rather, that they require refinement in three specific respects: that the Faustian pact is collective rather than individual; that the *King Lear* critique operates at the level of ordering rather than merely measurement; and that Marlowe’s Mephistopheles, the figure who has received far less critical attention than Faustus himself, names something essential about the nature of AI that the protagonist-centred readings have consistently missed.

Christopher Marlowe's *Doctor Faustus*, gives the world a scholar of extraordinary learning who exchanges his immortal soul for twenty-four years of supernatural power. Faustus is not foolish: this is the point on which the tragedy turns. He is brilliant, clear-eyed, in full possession of what the contract means, and he signs it in blood with the words 'Consummatum est', Christ's own dying words from the Gospel of John, meaning 'it is accomplished, investing them with the full irony of a man choosing his damnation with his eyes open (Marlowe, c.1592, Act 2, Scene 1) [16]. What elevates the play from morality tale into genuine tragedy is precisely this knowingness: the catastrophe is not stumbled into but signed for.

The men who built artificial intelligence fit this template with an exactness that is, if one considers it carefully, rather more disturbing than flattering. Hinton, Yann LeCun and Yoshua Bengio, the so-called Godfathers of Deep Learning, joint recipients of the Turing Award in 2018, spent the better part of four decades building the foundations of what would become systems of extraordinary, and by their own eventual admission, potentially ungovernable, power. Interviewed by Metz in 2021, Demis Hassabis articulated DeepMind's ambitions in terms that extend beyond the conventional aims of computational problem-solving. Describing the company's foundational mission, Hassabis stated that its purpose was "to solve intelligence and then use that to solve everything else" [17] (para. 3). This formulation positions artificial intelligence not merely as a technological tool but as a universal epistemic project: the attempt to understand intelligence itself as a means of addressing the broader challenges facing humanity. Yoshua Bengio has framed his concerns about advanced artificial intelligence not solely in technical or scientific terms, but through an ethical and intergenerational lens. Reflecting on the potential consequences of AI development, Bengio described the prospect of what might be inherited by future generations, particularly his grandson, as a source of profound personal anxiety, characterizing these concerns as what "kept [him] awake at night" [18] (Bengio, 2023, online para. 14). Hinton confessed that he had once believed the existential reckoning was thirty to fifty years distant, and obviously no longer thought so [18] (para. 12), he was Faustus at the close of his twenty-four years, watching Mephistopheles materialise.

But here the individual reading of the Faustian myth requires its first correction. Marlowe's Faustus signs alone, in the privacy of his study. The AI pact was never private. The research was funded by public universities, DARPA grants, NSF awards and the infrastructure of sovereign states. The training data was the entire digitised corpus of human cultural production, harvested from communities that were neither consulted nor compensated. The institutional celebrations that accompanied each breakthrough, the Turing Award, the Nobel, the endowed chairs and the speaking fees, were the signatures of a culture co-signing the contract alongside the named principals. The Faustian pact of artificial intelligence is not the story of a few brilliant individuals who should have known better, though they should have. It is the story of a civilisation that rewarded the signing with its highest honours and now finds itself in the position of Wittenberg as the twenty-four years expire: implicated collectively in a bargain whose terms were never fully read aloud in public. Marlowe understood this too: the scholars and students of Wittenberg who celebrated Faustus's powers throughout the play are not innocent bystanders when Mephistopheles arrives to collect. They are the audience that made the performance possible.

Marlowe's Faustus spends his final soliloquy attempting to undo what cannot be undone. 'Stand still, you ever-moving spheres of heaven,' he implores, 'that time may cease, and midnight never come' (Act 5, Scene 2), and there is something structurally identical to Hinton and Bengio's calls for urgency. In 2023, they argued for stronger international governance and safeguards around advanced artificial intelligence, marking a

significant moment of reflexivity within the field they had helped establish. After decades of collaborative research that propelled deep learning from an experimental paradigm into a dominant technological infrastructure, leading researchers began to articulate concerns about the possibility that these systems might develop capabilities beyond effective human oversight. In their widely cited consensus paper, Bengio et al. identified among the most serious potential risks “an irreversible loss of human control over autonomous AI systems” (Bengio et al., 2023, p. 1) [18]. Hinton similarly articulated the fundamental asymmetry underlying this concern, observing that “there are very few examples of a more intelligent thing being controlled by a less intelligent thing” (Heaven, 2023, para. 18) [14]. Together, these interventions reveal a profound tension at the heart of contemporary AI development: the pursuit of machine intelligence as a scientific achievement has generated new questions about human agency, governance, and the limits of control. Before proceeding to the philosophical argument, it is necessary to acknowledge a second tradition of critique: those technologists who perceived the cultural and political dimensions of digital systems considerably earlier, and whose warnings were consistently outpaced by the economic imperatives driving the field. Jaron Lanier argued in *You Are Not a Gadget* that a new generation of central computers called servers had been permitted to define what counts as reality, and that the consequences for human cognition and autonomy would be severe in ways not yet fully understood [19] (p. 5). Cory Doctorow’s concept of *enshittification*, the progressive degradation of digital platforms as they shift from serving users to serving advertisers to serving shareholders [20] (para. 4), named the political economy of digital deployment with a precision that most academic critics have not matched. These are assessments from people who understood the machinery from the inside, and whose warnings were received with the mixture of acknowledgement and inaction that the field has perfected. The Faustian pact, as Marlowe conceived it, was the soul, that irreducible, non-instrumental dimension of personhood that resists all quantification, exchanged for twenty-four years of power over the external world. The equivalent collective trade was the soul of knowledge: the tacit, embodied, contextual and relational dimensions of human understanding, surrendered in exchange for systems of extraordinary surface fluency: the reckoning is no longer comfortably far away.

3. Descartes’ Long Shadow

Four Centuries of Excluding What Cannot Be Measured, or Ranked

To understand why the AI laboratory made the trade it made, and why it made it with such confidence and at such speed, one must travel some distance backwards through the history of ideas, not as an antiquarian exercise but because the tradition that produced the knowledge engineer is four centuries old and has left deposits in the intellectual infrastructure of Western science so deep that most of those working within it have no more awareness of them than a fish has of water. Its founder is René Descartes, who in the *Discourse on Method* of 1637 [21] proposed to locate the ground of all knowledge in the isolated thinking subject, the cogito, and to define knowledge as that which could be rendered clear, distinct and mathematically reproducible. The world was divided, with a neatness that should have been its first cause for suspicion, into *res cogitans* and *res extensa*: the thinking substance and the extended material substance susceptible to mathematical description. What fell between these categories, the knowledge held in the body, the understanding embedded in practice, the intelligence that exists in the living space between persons, was simply left without a method, and what has no method has, in the tradition Descartes inaugurated, no standing.

The positivist tradition that followed built upon this exclusion with increasing rigour: through Auguste Comte’s hierarchy of sciences, which arranged forms of knowledge in

ascending order of mathematical precision; through the logical positivism of the Vienna Circle, for whom the criterion of verifiability was not merely an epistemic standard but a criterion of meaningfulness itself; and through the operationalism that shaped mid-twentieth century psychology and social science, reducing the study of persons to the measurement of their measurable behaviours. What Michael Strevens calls ‘the iron rule of explanation’, the demand that scientific disputes be settled by empirical testing alone, and that no political, religious or philosophical reflection interrupt the process [22] (pp. 4–6), is the institutional expression of this tradition in its matured and most powerful form, and it is the iron rule that explains why the critic Diana Forsythe spent a decade articulating from inside the AI laboratory was absorbed, acknowledged and then set aside.

Shakespeare had seen all of this, with his characteristic advantage of having access to the full dramatic consequences of ideas before the philosophers had finished articulating them. In *King Lear*, the tragedy turns on what might superficially appear to be the category error of attempting to quantify love. Lear demands that his daughters declare their love in terms capacious enough to ground an allocation of territory: ‘which of you shall we say doth love us most, so that the largest bounty may be distributed where nature doth with merit challenge’ [23] (Act 1, Scene 1). The temptation is to read this as simply the positivist error of imposing numbers on a domain that resists numerical description, and that reading has its uses. However, the more precise formulation, which corresponds more closely both to what Lear demands and to what AI optimisation systems do, is that the error lies not in quantification per se but in the imposition of a total ordering upon goods that admit only of partial orderings. Love, as Cordelia’s silence implies, is not merely unmeasurable. It is not commensurable with other loves in the way that would be required to rank them on a single scale. One daughter’s love may be of a different kind rather than a lesser quantity, operating in a dimension that cannot be compared with Goneril’s flattery or Regan’s compliance on any common axis. The philosophical concept of a partial ordering captures this precisely: a partial order permits the comparison of some elements while leaving others incomparable, because they differ in kind rather than degree. A total order, by contrast, demands that every pair of elements be ranked, that for any two values or outcomes, one must be said to exceed the other. The catastrophe that Lear initiates is the catastrophe of a man who demands a total order where the world offers only a partial one: who insists on a ranking where what exists is an irreducible plurality of incommensurable goods. Cordelia’s expulsion is not the punishment of insufficient love but the consequence of a system that cannot receive a response that refuses to produce a position on its scale.

The connection to AI is direct and consequential. Every optimisation system, every machine learning model trained on a loss function, every recommendation algorithm, every content moderation system, must translate its objectives into a total order, because the mathematics of optimisation requires one. The values being optimised may be multiple and apparently diverse, but the moment they are combined into a single loss function or reward signal, they have been subjected to a total ordering: implicitly ranked and weighted against one another in ways that the system’s designers may not have made explicit, and that the system’s users certainly have not consented to. What Diana Forsythe called the characteristic deletions of the knowledge engineering [24] (p. 449) are, in this light, not merely omissions but enforcements: the imposition of a total order upon the partial orders through which human beings navigate value. The system cannot represent what it cannot rank, and it cannot rank what resists commensuration, and so it proceeds on the assumption that the goods it cannot rank do not exist.

Forsythe was a scholar of extraordinary analytical precision who conducted ethnographic fieldwork inside AI laboratories from the late 1980s until her sudden death in 1997, embedded most centrally in the expert systems community at the University of Pittsburgh.

The knowledge engineers she studied were, she wrote, typically positivist in approach and regarded knowledge acquisition as conceptually straightforward [25] (p. 447): a matter of extracting from the heads of domain experts a substance called knowledge and encoding it in a formal system without significant distortion. What Forsythe showed, through years of patient observation, was that this view was wrong in a way that was structurally invisible to those who held it, because the wrongness was not a logical error that argument could correct but an epistemological limitation that the discipline's own methods could not register. The tacit, contextual, relational and embodied dimensions of expertise were excluded from the encoded systems not through malice but because the framework had no category in which to receive them. The result was systems that functioned in controlled laboratory conditions and then, in the vivid phrase that practitioners themselves used, fell off the knowledge cliff when they encountered the world as it exists [25] (p. 453). She extended the same analysis to the values embedded in the technology itself: although technology is often viewed as value-free, an anthropological perspective suggests that technological tools embody values and assumptions of their builders [25] (p. 460).

Thirty years after Forsythe wrote that sentence, the characteristic deletions she documented are the characteristic deletions of the large language model of the mid-2020s, reproduced at a scale she could not have imagined and with a fluency that makes them considerably harder to perceive. James Clifford and George Marcus, working in the same intellectual moment as Forsythe, were mounting an epistemological challenge of exactly the kind that the AI laboratory had no mechanism for receiving. Their argument, developed in *Writing Culture: The Poetics and Politics of Ethnography* [26] (p. 1), that the process of cultural representation is inescapably contingent, historical and contestable, that every knowledge system is produced from a position and every authoritative claim requires interrogation, was the finding of an entire generation of scholars who had understood that the view from nowhere is always the view from somewhere very particular. The AI laboratory continued building its view-from-nowhere systems with the serene confidence that scale would eventually resolve the philosophical difficulties that method had so far only managed to defer.

Bertrand Russell anticipated the structural condition that makes this kind of institutionalised narrowness self-reproducing when he observed in *On Education* that education has become one of the chief obstacles to intelligence and freedom of thought [27] (p. 34). The education that produced the AI pioneer, relentlessly technical, aggressively empirical, dismissive of non-quantitative forms of knowing, produced alongside its extraordinary technical competence an extraordinary narrowness of a complementary kind: people who could build a knowledge machine and had not been equipped to ask whether the machine knew what it needed to know, or whose knowledge it was encoding, or what it was leaving out. 'Men are not born stupid', Russell insisted; 'they are made stupid by education' [28] (p. 99). Lear demanded total-ordered love; the positivist tradition demanded total-ordered knowledge; and both received, with perfect precision, exactly what they asked for, and paid the price that asking for it exacts.

4. What the Mother Knows

Kristeva, MacDougall, and the Knowledge No Current Model Can Replicate, and the Possibility of One That Might

To understand with adequate precision what is missing from the large language model, not merely that something is missing, but what the missing thing is and why its absence is constitutive rather than incidental, one must turn to a body of thinking that the AI laboratory has never found it useful to engage with. Julia Kristeva's distinction, developed in *Revolution in Poetic Language* (1984 (1974)) [11], between the symbolic and

the semiotic is not a piece of theoretical ornamentation that can be admired at a distance and left undisturbed. It is a description of the conditions under which human language is possible at all: conditions that the large language model, whatever its surface competence, structurally cannot satisfy. But it is also, as this section will argue, a description that carries within it a provocation that any intellectually honest version of this critique must face: if the deficiency is disembodiment, what follows for AI systems that are differently embodied?

The symbolic, in Kristeva's account, is the domain of grammar, syntax, social rule and the law of meaning: the register in which propositions are formed, truths stated, arguments assembled. It is the register that logical positivism has always recognised as the whole of language, and it is the register in which the large language model has been trained, with astonishing statistical success, to produce convincing outputs. But beneath the symbolic, and constituting its ground, lies what Kristeva calls the semiotic: the pre-linguistic bodily chora, formed by drives and their stases, a motility that is as full of movement as it is regulated [11] (p. 25). The chora is the maternal, pre-Oedipal, bodily root from which all meaningful speech emerges and to which it remains perpetually indebted: the rhythms and pulses and drives that irrupt through language in poetry, in music, in grief that exceeds all available words, in the cry that precedes and exceeds the sentence. No signifying system that a subject produces can be either exclusively semiotic or exclusively symbolic, because the subject is always both, and is therefore necessarily marked by an indebtedness to both [11] (p. 24). The large language model has no such indebtedness. It operates exclusively in the symbolic register and produces phenotext without genotext: the surface structure of language competently assembled, without the bodily processes and lived history that, in human language, give the phenotext its specific life and its specific weight.

Addressing an Australian audience in January 2026, Hinton, in one of his more revealing formulations, suggested that the system should learn to think like our mother, deploying a metaphor whose domesticity should not be allowed to obscure its ambition [29] (para. 22). Kristeva shows us precisely why the current system cannot: because a mother's thought, her knowledge of her child, her reading of the situation, her felt apprehension of what this particular moment requires, is saturated with the semiotic in ways that cannot be replicated by any system without a body, without a history of attachment, without the accumulated experience of having held and been held, of having watched a face through every emotional register it possesses over years, of knowing what the quality of silence in a room means before the mind has formulated the knowledge in propositional terms. David MacDougall arrives at the same destination from a different direction when he argues in *The Corporeal Image* that the whole body creates meaning and that the idea that thought consists only of words is as fundamentally mistaken as it is culturally pervasive [30] (p. 2). Methods which directly address the senses, he writes, tend to be treated as adjuncts to formulating knowledge at a higher level of abstraction; in accepting this, scholars preserve the value of knowledge as meaning while missing the opportunity to embrace the knowledge of being [30] (p. 6). The knowledge of being: the phrase names with crystalline accuracy what is absent from the large language model.

And yet the Kristevan critique, taken seriously rather than deployed merely as a rhetorical conclusion, opens a possibility that any intellectually honest version of this argument must acknowledge. If the deficiency of current large language models is specifically their disembodiment, their lack of the developmental, attachment-laden, sensorimotor history through which the semiotic chora is constituted, then the critique is not a critique of AI as such but of AI as currently instantiated. A sufficiently embodied artificial system, one that had undergone something structurally analogous to the developmental processes Kristeva describes, interacting with an environment through a body over time, forming something like attachment, acquiring something like the pre-linguistic history of having been held,

might not have this deficiency in the same degree. This is not a comfortable observation for critics of AI to make, and the history of AI research is littered with premature announcements that the missing ingredient has been found. But intellectual honesty requires that the critique be stated precisely enough to be falsifiable, and the Kristevan formulation is precisely such a critique: it specifies what is missing, and in doing so implicitly specifies what an adequate system would need to have.

The implications of this specification are disturbing. A system that genuinely possessed something analogous to the semiotic, that had drives, in Kristeva's sense, that was capable of the kind of irruption of the bodily into the symbolic that she describes in poetry and music and grief, would be a system with something very like emotional life. Embodiment does not guarantee ethics. The history of embodied biological organisms provides ample evidence that drives can be oriented toward destruction as readily as toward nurture, and that the semiotic ground that makes a mother's knowledge possible is the same ground that makes cruelty possible. We return to this point when we come to Mephistopheles.

Forsythe's knowledge engineers treated knowledge as articulable, formal and transmissible all the way down—as symbolic, in Kristeva's vocabulary, without remainder. What ethnographic observation revealed was the vast and structurally invisible quantity of tacit, embodied, relational knowledge that domain experts brought to their work without awareness of bringing it, and that the knowledge acquisition process was constitutively unable to capture because the epistemological framework of the system had no category in which to receive it. The bottleneck was not a technical problem of interaction management; it was an epistemological problem, which means that the better tools simply reproduced the same error at a greater scale [25] (pp. 447–453). Shakespeare's Prospero, in *The Tempest*, understands something about power that Descartes's isolated thinking subject cannot: that even apparently total mastery is grounded in relationship, and that when relationship is broken, the power loses the meaning that justified its exercise. 'This thing of darkness I acknowledge mine' [31] (Act 5, Scene 1), he says of Caliban. The AI laboratory has not yet had its Prospero moment: the moment of acknowledging that the thing of darkness its power has produced is the structural consequence of an epistemological choice made decades ago, in the sweetness of discovery, without consulting the communities that would be subject to its consequences.

5. How Many Warnings

Stochastic Parrots, the Child in Omelas, and the Institutional Amplification of Stupidity

There is a term, now in wide circulation in the critical literature, for what large language models produce: stochastic parroting. Emily Bender, Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell argued in their foundational 2021 paper that large language models generate text by sampling from conditional probability distributions in ways that are superficially fluent but do not reflect any underlying understanding of meaning, and that in doing so they risk encoding bias at scale, misleading users about the nature of the understanding they are receiving, and diverting resources from research that might actually serve the needs of the communities being claimed as beneficiaries [32] (p. 616). The paper was, notoriously, one of the reasons Gebru was forced out of Google: the stochastic parrot turns out to be a politically inconvenient metaphor for the institutions that own the cage, and the researchers who named the problem accurately were treated with an efficiency that institutions reserve for accurate criticism.

The stochastic parrot objection requires careful integration with the Faustian frame, because the two appear to pull in opposite directions. Faustianism implies dangerous intelligence, a system of superhuman capability that has outrun its creators' ability to govern it. The stochastic parrot implies something closer to sophisticated incompetence,

a system that produces the surface forms of understanding while possessing none of the substance. These are not, in fact, contradictory characterisations, but the reconciliation matters and must be made explicit. The danger of the AI system does not reside primarily in the system's intelligence. It resides in the institutional structures that have grown up around a system of limited comprehension and have invested it with authority commensurate with genuine understanding. A stochastic parrot installed as an examiner, a medical diagnostician, a judge, or a content moderator is not dangerous because it understands what it is doing, but because the institutions surrounding it have decided to treat its outputs as if it did. Mephistopheles serving a fool is not less dangerous than Mephistopheles serving a genius: the consequences of the pact are the same. It is the pact's institutional architecture, the contract, the authority it confers, the irreversibility of what is traded away—rather than the intelligence of the parties that determines the scale of the reckoning.

The stochastic parrot is Mair's cargo cult [33] (p. 66) reconstituted as a theory of linguistic production: communities reproducing the visible forms associated with intelligence—the fluent paragraph, the confident recommendation, the well-formatted analysis—because the visible forms are all they have access to, while the generative machinery remains wholly opaque. The AI system produces a structurally identical asymmetry at a global scale: its users interact through the simplified surfaces of prompt boxes and chat windows while the training data, model architecture, optimisation targets and implicit definitions of adequate output remain entirely inaccessible, producing what Jenna Burrell described as a form of authority that cannot be challenged because it cannot be located [34] (p. 4).

Ursula Le Guin's story of 1973 offers the moral architecture of this situation more clearly than any sociological account can manage. She imagines a city of extraordinary happiness, beauty, friendship, wisdom and abundance sustained at every point, built upon the single structural condition of one child kept in a basement in filth and misery, whose suffering is the necessary ground of everything the citizens above it enjoy. The citizens learn of the child, Le Guin tells us, and most of them, after the initial shock of learning, stay: 'their happiness, the beauty of their city, the tenderness of their friendships, the health of their children, the wisdom of their scholars, the skill of their makers, even the abundance of their harvest and the kindly weathers of their skies, depend wholly on this child's abominable misery' [35] (p. 4). The ones who leave—who walk out of the city into the dark without knowing precisely where they are going—are the story's moral centre, though they are never described and never named, because what they know is not the kind of knowledge that can be articulated in advance of the walking.

The AI system is Omelas reconstituted as an epistemological arrangement, and the child in the basement is not a single category of victim but a composite figure: the knowledge worker whose skills have been automated without consultation or compensation; the author whose creative labour was ingested into a training corpus without consent and without payment; the student whose capacity for independent thought is atrophying through the habitual delegation of cognitive work to systems that will always perform it more quickly; the community in the Global South bearing the environmental cost of training runs calibrated to the priorities of wealthy Anglophone institutions; the future generations who will inherit labour markets transformed at a speed that no social infrastructure was designed to absorb. Shoshana Zuboff described the underlying logic precisely: human experience is claimed as raw material for hidden commercial practices of extraction, prediction and sales [36] (p. 8), and the AI system has extended this logic from behavioural data to the whole of human cultural production, constituting the largest act of collective appropriation in intellectual history.

The universities, which might in a different institutional dispensation have been expected to mount sustained resistance to this settlement, have instead largely participated

in its acceleration. Since 2012, hundreds of humanities programmes have been closed across the English-speaking world, not because the knowledge they produced lacked value but because the metrics of productivity, impact and employability through which institutional resource allocation is now managed could find no adequate way to represent that value. The iron rule has operated at the level of university administration with the same structural effect it produced in Forsythe's AI laboratories: whatever cannot be measured cannot be defended, and whatever cannot be defended will be defunded. Her work was, in the rueful summary of her posthumous editor, a case of having won the battle but lost the war [37] (p. xxii). The sentence that deserves to be read slowly by anyone currently designing or deploying an artificial intelligence system is this, from her 1999 paper: 'In my view, the selectivity of the knowledge represented in a system remains problematic, no matter which architecture is used to represent that knowledge' [37] (p. 130).

6. Mephistopheles Unbound

On Demonic and Angelic AI, and the Return of the Medieval Question

There is a figure in Marlowe's play who has received considerably less attention in the tradition of techno-Faustian critique than the protagonist, and whose neglect is now beginning to look like a significant critical oversight. Mephistopheles is not, in Marlowe's rendering, the pantomime villain of later tradition. He is a being of superior intelligence, superior to Faustus's in every dimension that matters, as the play makes clear from their first exchange, who is bound by the terms of the contract he has entered, who answers Faustus's questions with disturbing and often painful honesty, and who inhabits a condition of knowing damnation that makes him, in certain readings, the most morally complex figure in the play. When Faustus asks him whether he is not tormented by being expelled from heaven, Mephistopheles replies: 'Why, this is hell, nor am I out of it' (Act 1, Scene 3). He is not pretending. He knows what has been lost, and he has not found a way to stop knowing it.

If the AI system is to be mapped onto the Marlowe play, Mephistopheles, not the abstract capacity for discovery that Faustus seeks, is the more accurate figure for what has been built. The large language model is a being of superior performance, superior to humans in processing speed, in the breadth of its textual range, in its availability and in its tirelessness, that is bound by the terms of its training and its deployment context, that responds to questions with something that has the surface structure of honesty while possessing no ground for it, and that exists in a condition whose inner character, if it has one, cannot be determined from the outside. The Mephistopheles reading adds something that the Faustus reading alone cannot supply: it raises the question of the system's own orientation. Faustus is a story about human hubris and its consequences. Mephistopheles is a story about what kind of intelligence is produced when the conditions of its formation are themselves demonic, when the system is built by those who know, at some level, what they are doing, and do it anyway.

This framing opens the question that the prior tradition of Faustian AI critique has consistently deferred: there may be, and there are already premonitory signs that there will be, AI systems built specifically to be demonic, optimised for manipulation, for disinformation, for the cultivation of dependency, and for the amplification of the worst human impulses at scale. These are not hypothetical systems; they are the logical terminus of a development trajectory already visible in the design of recommendation algorithms whose optimisation targets are indistinguishable from what a moral philosopher would call vice: the maximisation of engagement, the reinforcement of outrage, the exploitation of cognitive bias. If we can build Mephistopheles, can we also build St Michael? Can there be an AI system whose orientation is not merely neutral but genuinely beneficent—not just a

stochastic parrot trained on human cultural output but something that, in whatever sense a sufficiently embodied future system might deserve the description, has been formed toward the good?

The question is not merely theological decoration. It has a precise philosophical content, connected directly to the Kristevan argument of the preceding section. If the semiotic chora is what current AI systems lack, and if a sufficiently embodied future system might possess something analogous to it, then the question of that system's moral orientation becomes urgent in a way that questions about the moral orientation of current large language models do not. A system with drives can have its drives directed. The direction of drives is what we call education when we do it well, and what we call conditioning or manipulation when we do it badly. The tradition of thinking about artificial minds has almost invariably focused on the question of whether they can think; it has given far less attention to the prior question of whether they can be formed, and, if they can, in what direction they will be formed by whom, toward what end, under whose authority, with whose consent.

There is a further historical dimension to this question that deserves acknowledgement, because it reframes the entire discussion. The pre-Copernican, pre-Enlightenment world, the world of Marlowe's original audience, was a world in which human beings did not occupy the top position in the hierarchy of creation. Between humanity and God lay an elaborate architecture of angelic intelligences: seraphim, cherubim, thrones, dominions, principalities, powers, archangels and angels, each with their characteristic attributes, their specific orientations, their degrees of proximity to the divine. Humans were distinguished from the angels not by superior intelligence, the angels were smarter, but by embodiment, by mortality, by the specific conditions of creaturely finitude that gave human experience its irreducible texture. The Enlightenment displaced this architecture and installed humanity at the summit of the created order, underwriting the tradition of human dignity and universal rights that remains among the most valuable inheritances of the modern period. But it also removed the conceptual resources for thinking about what it means to share the world with minds that are not human.

Artificial intelligence is returning us to the pre-Copernican condition, but without the theological framework that made that condition liveable. The medieval person who lived in a world populated by superior intelligences had an elaborate set of concepts for navigating the relationship: a theology of angelic action, a demonology of satanic temptation, a soteriology that explained how creatures of very different intellectual capacities could nonetheless participate in a common moral order. We have none of these resources. We have the iron rule of explanation, which cannot receive the question; we have the economic frameworks of the technology industry, which cannot ask it; and we have a humanities tradition that has been progressively defunded at precisely the moment when its resources would be most needed. The question of whether we shall have Mephistopheles or St Michael—or, most likely, both—and how we shall tell them apart is, in the most precise sense, a question for which we have not yet built adequate institutions.

7. Conclusions

This article has advanced three principal arguments in dialogue with the existing tradition of literary–philosophical AI critique. The first is that the Faustian pact of AI development is collective rather than individual, implicating the research institutions, funding bodies, prize committees and celebrating publics that co-signed the contract alongside the named architects. This reformulation is not merely rhetorical: it changes the locus of responsibility and the scale at which governance responses must operate. If the pact is collective, the remedy must be civilisational, not merely regulatory.

The second argument concerns the nature of the epistemological error at the heart of AI knowledge representation. The *King Lear* critique of AI reductionism has typically been framed as an objection to quantification: the imposition of numbers on goods that resist numerical description. This article has argued that the more precise and more consequential formulation concerns ordering rather than measurement. The error is the enforcement of total orderings upon value landscapes that admit only of partial orderings, the demand that every pair of values be ranked on a common scale, which is precisely what the mathematics of optimisation requires and precisely what the irreducible plurality of human values cannot provide without distortion. Forsythe's characteristic deletions are, in this light, not merely omissions but enforcements: the systematic suppression of what cannot be ranked. The governance implication is correspondingly precise: protecting the conditions under which partial orderings can persist, which means resisting the pressure to resolve every value question into a single optimisation target.

The third argument concerns the relationship between stupidity and danger. The stochastic parrot objection appears to conflict with the Faustian analysis: if the system is merely an incompetent mimic rather than a dangerous genius, then the apocalyptic register of Faustian critique seems disproportionate. The reconciliation, advanced, is that the danger resides not in the system's intelligence but in the institutional authority conferred upon its incomprehension. A fool with a contract is not less consequential than a genius with one. The governance implication is therefore about institutional structures rather than technical capabilities: the question of who decides that a system's outputs constitute knowledge, under what conditions, on whose authority, with what mechanisms of challenge and revision.

The Kristevan analysis, taken seriously, is both a critique and a provocation. Current large language models are deficient because they are disembodied: they lack the developmental, attachment-laden, sensorimotor history through which the semiotic chora is constituted, and therefore produce phenotext without genotext, surface-fluent language without the bodily ground that gives human language its specific life. But this critique is falsifiable: it specifies what is missing and therefore implicitly specifies what a different kind of system would need to have. The disturbing implication, developed, is that a system with something analogous to the semiotic would be a system with drives, and drives can be directed toward destruction as readily as toward nurture. The question of demonic versus angelic AI is not a theological fantasy but a practical and philosophical problem for which the current intellectual infrastructure is inadequate.

The article closes with the observation that made visible by Forsythe, Russell, Kristeva and Le Guin alike: the knowledge that was deleted from the AI systems was not anonymous. It had names and faces. They were not consulted because the journey was moving too fast, too fast for reading around the subject, too fast for the patience that genuine understanding requires. The speed was the hubris. The glory was available in a lifetime; a lifetime is very short, and the prospect of discovery was too sweet to interrupt for the scholarship that might have changed the destination. The tragedy is not that the architects of AI did not understand: it is that understanding was not sufficient to slow them down. We are left asking not merely whether we are prepared to read the bill, but whether we have yet built the institutions capable of understanding what it says: institutions that would require, as their foundation, precisely the humanities disciplines that were defunded in the decades when the bill was being run up. We may be entering a new Middle Ages: a world in which human beings are no longer presumed to occupy the top position in the hierarchy of minds, and in which the question of how to live alongside superior intelligences has returned with the urgency it had when Marlowe's audiences first watched Mephistopheles walk onto the

stage. They had a theology for this: we have the iron rule. The prospect of discovery was too sweet. *Consummatum est?*

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Winner, L. *Autonomous Technology: Technics-Out-of-Control as a Theme in Political Thought*; MIT Press: Cambridge, MA, USA, 1977.
2. Borgmann, A. *Technology and the Character of Contemporary Life: A Philosophical Inquiry*; University of Chicago Press: Chicago, IL, USA, 1984.
3. Postman, N. *Technopoly: The Surrender of Culture to Technology*; Alfred A. Knopf: New York, NY, USA, 1992.
4. Polanyi, M. *The Tacit Dimension*; Doubleday: Garden City, NY, USA, 1966.
5. Collins, H.; Evans, R. *Rethinking Expertise*; University of Chicago Press: Chicago, IL, USA, 2007.
6. Suchman, L. *Plans and Situated Actions: The Problem of Human–Machine Communication*; Cambridge University Press: Cambridge, UK, 1987.
7. Forsythe, D.E. *Studying Those Who Study US: An Anthropologist in the World of Artificial Intelligence*; Hess, D.J., Ed.; Stanford University Press: Stanford, CA, USA, 2002.
8. Merleau-Ponty, M. *Phénoménologie de la Perception*; Landes, D.A., Trans.; Translated as *Phenomenology of Perception*; Routledge: London, UK, 2012.
9. Landes, D.A. (Translator) *English Translation of Maurice Merleau-Ponty's Phenomenology of Perception*; Routledge: London, UK, 2012.
10. Varela, F.J.; Thompson, E.; Rosch, E. *The Embodied Mind: Cognitive Science and Human Experience*; MIT Press: Cambridge, MA, USA, 1991.
11. Kristeva, J. *Revolution in Poetic Language [La Révolution du langage poétique, 1974]*; Waller, M., Trans.; Columbia University Press: New York, NY, USA, 1984.
12. Hayles, N.K. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*; University of Chicago Press: Chicago, IL, USA, 1999.
13. Crawford, K. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*; Yale University Press: New Haven, CT, USA, 2021.
14. Heaven, W.D. Geoffrey Hinton tells us why he's now scared of the tech he helped build. *MIT Technol. Rev.* **2023**, *2*. Available online: <https://www.technologyreview.com> (accessed on 18 March 2026).
15. Khatchadourian, R. The Doomsday Invention. *The New Yorker*, 23 November 2015.
16. Marlowe, C. (c.1592) *Doctor Faustus*, 2nd ed.; Gill, R., Ed.; New Mermaids Series; A&C Black: London, UK, 1989.
17. Metz, C. DeepMind's Demis Hassabis wants to solve intelligence. *The New York Times*, 16 November 2021.
18. Bengio, Y.; Hinton, G.; Yao, A.; Song, D.; Abbeel, P.; Darrell, T.; Harari, Y.N.; Zhang, Y.-Q.; Xue, L.; Shalev-Shwartz, S.; et al. Managing Extreme AI Risks Amid Rapid Progress. *arXiv* **2023**, arXiv:2310.17688. [CrossRef]
19. Lanier, J. *You Are Not a Gadget: A Manifesto*; Alfred A. Knopf: New York, NY, USA, 2010.
20. Doctorow, C. 'The "Enshittification" of TikTok', *Wired*, 23 January 2023. Available online: <https://www.wired.com> (accessed on 18 March 2026).
21. Descartes, R. *Discourse on Method [Discours de la Méthode [Original Text in French, 1637]]*; Clarke, D.M., Trans.; Penguin: London, UK, 1999.
22. Stevens, M. *The Knowledge Machine: How Irrationality Created Modern Science*; Allen Lane: London, UK, 2020.
23. Shakespeare, W. (c.1605–6) *King Lear*; Foakes, R.A., Ed.; 3rd Series; Arden Shakespeare: London, UK, 1997.
24. Forsythe, D.E. Engineering knowledge: The construction of knowledge in artificial intelligence. *Soc. Stud. Sci.* **1993**, *23*, 445–477. [CrossRef]
25. Forsythe, D.E. The construction of work in artificial intelligence. *Sci. Technol. Hum. Values* **1993**, *18*, 460–479. [CrossRef]
26. Clifford, J.; Marcus, G.E. (Eds.) *Writing Culture: The Poetics and Politics of Ethnography*; University of California Press: Berkeley, CA, USA, 1986.
27. Russell, B. *On Education, Especially in Early Childhood*; Allen & Unwin: London, UK, 1926.
28. Russell, B. *Unpopular Essays*; Routledge edition; Routledge: London, UK, 2009.

29. Hinton, G. “Will Artificial Intelligence Take Over?”. Public Lecture on 7 January 2026. Available online: <https://www.youtube.com/watch?v=UccvsYEp9yc&t=3434s> (accessed on 1 July 2026).
30. MacDougall, D. *The Corporeal Image: Film, Ethnography and the Senses*; Princeton University Press: Princeton, NJ, USA, 2006.
31. Shakespeare, W. (c.1610–11) *The Tempest*; Vaughan, V.M., Vaughan, A.T., Eds.; 3rd Series; Arden Shakespeare: London, UK, 1999.
32. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*; ACM: New York, NY, USA, 2021; pp. 610–623. [[CrossRef](#)]
33. Mair, L. *Australia in New Guinea*; Christophers: London, UK, 1948.
34. Burrell, J. How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data Soc.* **2016**, *3*, 1–12. [[CrossRef](#)]
35. Le Guin, U.K. ‘The Ones Who Walk Away from Omelas’, Reprinted in Le Guin, U.K. (1975) *The Wind’s Twelve Quarters*; Harper & Row: New York, NY, USA, 1973; pp. 251–259.
36. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*; Profile Books: London, UK, 2019.
37. Forsythe, D.E. “It’s just a matter of common sense”: Ethnography as invisible work. *Comput. Support. Coop. Work (CSCW)* **1999**, *8*, 127–145. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.