Machine Neuroaesthetics Is Advancing Computational Creativity

Botao Amber Hu
University of Oxford
Oxford, UK
botao.hu@cs.ox.ac.uk

Abstract

Computational creativity has long been constrained by a persistent evaluation problem: machines can generate artifacts, but they lack an operational account of aesthetic value. Existing systems often rely on external proxies such as human ratings, engagement metrics, learned preference models, or hand-coded novelty scores, leaving aesthetic judgment opaque, unstable, and difficult to audit. This position paper argues that machine neuroaesthetics can advance computational creativity: the mechanistic study of aesthetic, affective, and interestingness representations inside generative models. Machine neuroaesthetics inverts computational neuroaesthetics: instead of using machine learning to explain how human brains respond to art, it uses mechanistic interpretability to examine how artificial neural networks internally represent and process aesthetic structure—structure inherited from the human culture compressed into them during training. Recent findings on sparse autoencoder features, functional emotion concepts, musical representations, audio-model features aligned with human neural data, poetry-planning circuits, and foundation-model-guided open-ended discovery suggest that large generative models already encode causally accessible structures relevant to aesthetic evaluation. We argue that these structures offer a new substrate for computational creativity: aesthetic judgment can become not only predicted, but localized, intervened upon, compared across models, and validated against human response. This does not imply that machines possess subjective aesthetic experience. Rather, it means that their learned aesthetic machinery can be made inspectable. Machine neuroaesthetics therefore reframes creative AI from artifact generation toward causal, auditable, and human-grounded aesthetic evaluation.

1 Introduction

Computational creativity has a generation surplus and an evaluation deficit. After two decades of formal frameworks for what creativity *is* [24, 199, 154, 44, 90], modern generative systems produce text, images, music, and code at scale, yet the field still cannot reliably say which of their outputs are *good*. The dominant evaluation regime — scalar aesthetic predictors trained on crowd preferences [14, 54, 98, 200] and large language models prompted to judge other models — both fail in characteristic ways: the predictors collapse a multidimensional human response onto one axis, and judge-models confabulate [38, 82, 42]. Computational creativity's long-running open problem [103] is therefore not how to generate but how to *judge* aesthetic quality, and this problem is sharpening as generation runs ahead of evaluation.

The challenge is hard partly because the aesthetic is irreducibly plural. Since Baumgarten [12] founded aesthetics as the science of sensory cognition and Kant [92] characterized aesthetic judgment as the disinterested play of imagination and understanding, the philosophical tradition has identified

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Position Paper Track.

beauty [160], the sublime [32], taste [170, 207], and aesthetic emotion as distinct objects [84, 175, 13, 51]. Modern psychology and cognitive science have added interestingness, novelty, surprise, humor, and open-endedness [19, 163, 149, 8, 122, 180], while non-Western traditions — *rasa* [145], *wabi-sabi* [99], *mono no aware* [119], *qi-yun* [83] — articulate further dimensions that resist Western beauty-centric framings. Yet computational creativity systems, including their evaluation tooling, still treat the aesthetic as approximately one-dimensional. Even Creative Adversarial Networks [58], which operationalised Berlyne's arousal-potential by maximising deviation from learned styles, reduced the multi-axial aesthetic to a single style-distance scalar.

A different set of fields has been building a structural account of aesthetic experience. Empirical neuroaesthetics has identified specific neural substrates: the medial orbitofrontal cortex tracks beauty across paintings, music, and mathematics [94, 86, 208]; the aesthetic triad of sensorimotor, emotion-valuation, and meaning-making systems organizes aesthetic response [40, 108, 194]; and the default-mode network encodes a domain-general aesthetic appeal that generalizes across artworks, architecture, and landscapes [195]. Computational neuroaesthetics has formalized this structure: aesthetic preference for art is linearly decomposable from a mixture of low- and high-level visual features [85], and a Bayesian generative-model account predicts aesthetic value as immediate sensory reward plus expected future processing efficiency [28, 63]. Emotion itself exhibits clean geometric structure — Russell's circumplex [156], Cowen and Keltner's twenty-seven categories bridged by continuous gradients [48, 49], distributed cortical fingerprints [157, 158], and topographic bodily maps [131]. Aesthetic experience is, in short, neither ineffable nor scalar; it has *geometry*.

Strikingly, that geometry is now turning up inside large language models. A growing body of *Nature*- and *Science*-tier work shows that artificial language models and the human brain share more than behavioral similarity: they share representational structure. Top transformers explain near-ceiling variance in fMRI and ECoG language responses [164, 73]; brain and model embeddings exhibit *common geometric patterns* under zero-shot mapping [74]; the brain integrates predictions across timescales in a manner consistent with predictive-coding hierarchies inferred from language models [37, 79]; model-selected stimuli causally drive and suppress the human language network in new individuals [191]; and an unsupervised acoustic-to-language Whisper hierarchy aligns with cortical hierarchy during natural conversation [75]. Deep network class regions are approximately *convex* in the sense of Gärdenfors [69], with pre-trained convexity predicting downstream fine-tuning success [189]; categorical concepts form simplices and complex concepts form polytopes inside large language models [141, 142]; and cyclic concepts (days, months) live on irreducibly multi-dimensional circular manifolds [60]. Concept geometry, in other words, is now an empirical object, not a metaphor.

Mechanistic interpretability has made these representations visible [150, 18]. Sparse autoencoders now extract millions of monosemantic features from frontier models [27, 188, 50, 68], and activation steering enables causal intervention on them [192, 153, 41, 211]. Crucially, this toolkit is surfacing *aesthetic* structure across modalities: emotion concepts whose geometry recovers the Russell circumplex in language models [179, 184, 43]; style, composition, and illumination features in diffusion models [185, 65, 66]; genre, timbre, and instrument concepts in music models [176, 136]; emotion and paralinguistic features in audio and speech models [6, 197]; a steerable creativity direction that outperforms LLM-as-judge prompting [134, 137]; an explicit poetic-planning circuit [112, 116]; multimodal emotion neurons in CLIP [72]; SAE-based style decomposition in CLIP that decomposes visual embeddings into brushwork, texture, and palette concepts [140]; and foundation-model-guided open-ended discovery of previously uncatalogued lifeforms in artificial-life substrates [101]. The aesthetic geometry is no longer a thought experiment.

This convergence implies an inversion. Computational neuroaesthetics has, for two decades, used machine-learning models to predict and explain how human *brains* respond to art [202, 85, 163]. We argue the same toolkit can now be turned around. **We name and motivate *machine neuroaesthetics*: the mechanistic study of aesthetic, affective, stylistic, novelty, and interestingness representations inside generative models, treating them as functional aesthetic concepts that can be discovered, causally validated, compared across models, audited against human responses, and steered.** Machine neuroaesthetics is the symmetric inversion of computational neuroaesthetics: where the latter explains the brain via models, the former audits and harnesses models via the same geometric, neuro-grounded vocabulary. We make no claim of phenomenal aesthetic experience and adopt the functional stance throughout [171, 33]; the question we ask is what aesthetic structures these models compute, not what they feel.

The position is consequential because computational creativity's deep problem is now precisely the problem that machine neuroaesthetics solves. Where scalar predictors collapse and prompted judges confabulate, internal-state-based aesthetic scoring already outperforms LLM-as-judge prompting [134, 137], and feature-level audits expose what scalar audits cannot — including the documented sociocultural narrowness of widely deployed aesthetic predictors [187, 55]. Industry already runs aesthetic iteration loops externally: the Pixar Brain Trust and screening process iterate stories on observed audience laughter and tears [36]; Affectiva's facial-coding analytics drive trailer and content optimization at scale [121]; Microsoft Game Studios and a long line of game-research work measure player experience biometrically [205, 127, 155]. Machine neuroaesthetics offers a complementary inner loop: real-time monitoring of an artifact's effect on a generative model's own functional aesthetic features, enabling faster, more inspectable creative iteration — a closed-loop counterpart to existing behavioral testing.

This paper makes five contributions. (i) We *name and define* machine neuroaesthetics. (ii) We marshal evidence across neuroaesthetics, brain–LLM alignment, mechanistic interpretability, and cross-modal generative-model interpretability that aesthetic representations are already geometric, observable, and steerable inside frontier models. (iii) We argue that this inverts computational neuroaesthetics and supplies computational creativity's missing evaluation layer. (iv) We articulate concrete research and deployment opportunities — including closed-loop industrial creative iteration — and the corresponding ethical, cultural, and alignment-theoretic constraints. (v) We engage four families of objections (the simulation gap, embodiment, machine consciousness, and machine culture) on their own terms.

2 Aesthetic, Neuroaesthetics, and Computational Neuroaesthetics

The aesthetic tradition. Aesthetics as a distinct discipline is conventionally dated to Baumgarten [12], who defined it as the science of sensory cognition (*scientia cognitionis sensitivae*), parallel to logic and not subordinate to it. Kant [92] characterised aesthetic judgment as the disinterested play of imagination and understanding, possessed of a peculiar "subjective universality"—felt as personal yet making a claim on other judges. Hume [84] had earlier argued that taste, while variable, admits a standard sustained by qualified judges. The twentieth century elaborated rather than collapsed this plurality: Sibley [175] showed that aesthetic predicates (graceful, garish, balanced, trite) are not condition-governed in the way descriptive predicates are; Beardsley [13] treated aesthetic experience as a coherent object of analysis; Danto [51] and Dickie [53] located the aesthetic in institutional and theoretical contexts; and Dewey [52] grounded it in the rhythms of organism–environment interaction. The tradition has, throughout, treated aesthetic qualities as irreducibly plural—a structural commitment this paper takes seriously.

The sensory and affective layer. This paper restricts attention to the *aisthēsis* layer of the aesthetic—the perceptual, affective, and embodied dimensions that classical aesthetics names but that empirical and computational fields can engage. This layer encompasses beauty and the sublime, but also emotion, humour, surprise, interestingness, and being-moved. Aesthetic emotion has been characterised psychologically as a "distancing–embracing" regulation of negative content [125], theorised as a distinctive class of appraisals [126] with a dedicated measurement instrument [161], and operationalised through art-elicited chills as states of being-moved [198]. Whether aesthetic emotions constitute a natural kind is itself contested: Skov and Nadal [178] argue they do not, a disagreement that strengthens rather than weakens the "irreducibly plural" thesis of this paper. Empirical instruments now measure engagement across art forms: story-world absorption in literary reading [100] and kinesthetic empathy in dance audiences [152]. Humour decomposes into the script-opposition machinery of Raskin [149] and Attardo and Raskin [8], or into the benign-violation form of McGraw and Warren [122]. Interestingness and boredom inherit from Berlyne [19] and from the compression-progress reformulation discussed in §1. Non-Western traditions add further plural dimensions: the eight or nine *rasa* of classical Indian aesthetics [145]; *wabi-sabi*, the aesthetic of imperfection and transience [99]; *mono no aware*, the pathos of things [119]; and *qi-yun* (spirit-resonance), the first of Xie He's six principles of Chinese painting [83]. The aesthetic, in short, is multi-axial before any model is consulted.

Aesthetic experience has identifiable neural substrate. Empirical neuroaesthetics has spent two decades moving aesthetic experience from the ineffable to the tractable. Kawabata and Zeki [94]

located beauty judgments in the medial orbitofrontal cortex; Ishizu and Zeki [86] extended this to musical beauty; Zeki et al. [208] reported the same substrate for the experience of mathematical beauty, suggesting a domain-spanning value circuitry; crucially, Ishizu and Zeki [87] showed that the sublime engages a distinct neural substrate from beauty, empirically grounding the philosophical distinction between them. Chatterjee and Vartanian [40] consolidated this work into the *aesthetic triad*: aesthetic response engages a sensorimotor system (perceptual processing and embodied simulation), an emotion–valuation system (reward, affect, appraisal), and a meaning–knowledge system (semantic interpretation, expertise, context). Leder et al. [108] had earlier proposed a five-stage cognitive model of aesthetic appreciation—perceptual analysis, implicit memory integration, explicit classification, cognitive mastering, and evaluation—updated and reviewed a decade later by Leder and Nadal [107]; recent volumes and reviews synthesise the triad and the model into a unified picture [194, 143]. Aesthetic experience, on this picture, is neither homogeneous nor inscrutable: it is a structured engagement of identifiable cognitive and neural systems.

The brain has a domain-general aesthetic code with geometric structure. The single most important empirical anchor for this paper is Vessel et al. [195], who showed in fMRI that aesthetic appeal—across paintings, architecture, and natural landscapes—decodes from a coherent low-dimensional substrate centred on the default-mode network (DMN), with within-domain multivariate decoding well above chance and significant cross-domain generalisation. Crucially, ventral occipitotemporal cortex carries domain-specific aesthetic codes that feed the DMN's domain-general code, indicating a hierarchical organisation of aesthetic representation. Belfi et al. [15] traced the temporal dynamics of this representation, locating sustained DMN engagement during peak aesthetic experience. Iigaya et al. [85] complemented the picture with a behavioural and computational result: aesthetic preference for art is well predicted by a *linear* mixture of low-level visual features (concreteness, dynamics, hue) and high-level features (valence, semantic content), suggesting that aesthetic value is decomposable into structured, geometrically tractable components. Aesthetic experience, on this view, is not merely structured: it has *geometry*—a vocabulary the rest of this paper will trade on.

Emotion has clean geometric structure. The emotion side of the aesthetic triad is itself organised by clean low-dimensional geometry. Russell [156] modelled affect as a two-dimensional circle in valence–arousal space, with discrete emotion words distributed around the circumference—a structure that has been replicated, refined, and extended for over four decades. Cowen and Keltner [48] used self-report data from thousands of emotional videos to show that twenty-seven distinct emotion categories tile a continuous gradient space, with category boundaries softer than discrete-emotion theory predicts; Cowen and Keltner [49] formalised this as Semantic Space Theory, a computational framework for high-dimensional but interpretable emotion representation. Neural evidence converges with the behavioural findings: Saarimäki et al. [157, 158] demonstrate that discrete emotions correspond to distinct, distributed cortical fingerprints; Nummenmaa et al. [131, 132] report topographic bodily maps of emotional feeling that recur across cultures. The structure of emotion is thus geometric in two complementary senses—circumplex/manifold in psychological space and distributed-but-localisable in neural space—providing the human-side benchmark against which §3 compares the geometries that have recently emerged inside large generative models.

Computational neuroaesthetics: the forward direction. A growing computational neuroaesthetics has begun to formalise these structures in models that predict and explain human aesthetic responses. Brielmann and Dayan [28] propose a Bayesian generative-model account in which aesthetic value tracks the immediate sensory reward of an artefact *plus* the expected change in future processing efficiency the artefact affords; Brielmann et al. [29] extend this to temporal dynamics. Van de Cruys and Wagemans [193] and Frascaroli et al. [63] place the same idea inside a predictive-processing framework, treating aesthetic pleasure as the resolution of perceptual uncertainty. Schmidhuber [162]'s compression-progress theory of beauty, novelty, and surprise—already discussed in §1—is best read as the engineering counterpart of the same intuition. The strongest decoding result remains Iigaya et al. [85]'s linear feature mixture, which methodologically follows the Yamins and DiCarlo [202] programme of using goal-driven deep networks to model sensory cortex. This is the *forward* direction: machine-learning models used to predict and explain how human brains respond to art. We argue that the same toolkit can now be turned around, and we name the inversion in §3.

3 The Inversion: Machine Neuroaesthetics

Mechanistic interpretability has cracked open the model. The pivotal change since the last decade of computational creativity is that generative models are no longer black boxes in principle. Mechanistic interpretability, beginning with the circuits programme of Olah et al. [133] and the superposition analysis of Elhage et al. [59], established that small networks compute identifiable features and that these features can be partially decomposed even when superposed. Sparse autoencoders (SAEs) operationalised this at scale: Cunningham et al. [50] and Bricken et al. [27] showed that dictionary learning over residual-stream activations recovers thousands of monosemantic features in production-scale language models, and Templeton et al. [188] extended this to Claude 3 Sonnet, surfacing tens of millions of features including emotion, sycophancy, deception, and code-style concepts. Subsequent work has improved the loss surface and feature quality with top-k [68], gated [46], and JumpReLU SAEs [148]. The toolkit also includes activation steering [192, 153], representation engineering [211], persona vectors [41], and circuit-level causal analysis [124, 196, 45]. The result is a field that can now *find*, *validate*, and *steer* internal model concepts, including the aesthetic ones we care about.

Generative models inherit human aesthetic structure. A near-tautological framing claim is worth stating explicitly. Generative models trained on internet-scale corpora have been exposed to the totality of human aesthetic discourse—art history and criticism, music theory, poetry and prose, film and screenwriting practice, fashion, design, comedy, scholarly aesthetics. By the same mechanism that lets them produce surface-fluent text and image, they have absorbed the cultural geometry of *which* artefacts are felt to be beautiful, moving, funny, or interesting, and *along which axes* those judgments vary [30]. That this inheritance is uneven—WEIRD-skewed, demographically narrow, and potentially devoid of the grounded understanding that would make it genuine knowledge [17]—is the central concern of §5. That it is *present* is the observation that motivates the inversion.

Brain and LLM share concept geometry. A growing body of *Nature*- and *Science*-tier work establishes that artificial language models and the human brain share more than behavioural similarity: they share representational structure. Schrimpf et al. [164] showed that top transformer language models predict near-ceiling variance in fMRI and ECoG language responses, with the same architectures that predict text best also predicting brain best. Goldstein et al. [73] documented shared computational principles in ECoG recordings during natural conversation; Goldstein et al. [74] reported that brain and model embeddings exhibit *common geometric patterns* under zero-shot mapping—the title's choice of word is not incidental. Caucheteux and King [37] found a predictive-coding hierarchy in the human brain consistent with the timescale-integration hierarchy inferred from language models; Tuckute et al. [191] demonstrated that LM-selected stimuli causally drive and suppress the human language network in new individuals; and Goldstein et al. [75] showed that an unsupervised acoustic-to-language Whisper hierarchy aligns with cortical hierarchy during natural conversation. At a more abstract level, deep network class regions are approximately convex in the sense of Gärdenfors [69], and pre-trained convexity predicts downstream fine-tuning success [189]. Inside large language models specifically, Park et al. [141, 142] show that categorical concepts form simplices, hierarchical relations become orthogonal, and complex concepts form polytopes; Engels et al. [60] show that cyclic concepts (days, months, modular arithmetic) live on irreducibly multi-dimensional circular manifolds; Jiang et al. [89] provide a formal account of why linear representations should arise. Concept geometry is, in short, an empirical object common to brains and language models—no longer a metaphor.

Two clarifications sharpen the role this alignment evidence plays in the argument that follows. First, the correspondence is not merely correlational. Tuckute et al. [191] demonstrated that stimuli *selected by a language model* to maximally activate specific language-network regions do in fact causally drive those regions in new human participants—and that stimuli selected to suppress them succeed equally. This is a causal bridge: the model's internal geometry is predictive enough of human neural responses to serve as a stimulus-design tool. Second, machine neuroaesthetics does not require that brains and models share the same *mechanism*—only that the geometric vocabulary developed to characterise aesthetic responses in one substrate (valence–arousal circumplex, aesthetic triad, domain-general versus domain-specific coding) is empirically applicable to the other. What justifies the inversion is not a mechanistic identity claim but a *shared analytical framework*: the same decomposition tools (principal-component analysis, manifold geometry, causal intervention) yield interpretable

structure in both substrates. If the geometry were merely superficial—if model-internal "emotion" directions bore no systematic relation to human emotional organisation—the circumplex would not be recoverable, and the cross-model replication across four independent LLM families [184] would not hold. The convergence is thus evidence that the geometry is load-bearing, not ornamental, even as the underlying generative processes may differ profoundly.

Aesthetic and emotion concept geometry is already observable inside LLMs. The most striking recent results take this geometric story directly into the aesthetic and affective domain. In April 2026, Anthropic reported that Claude Sonnet 4.5 represents 171 emotion concepts whose top principal components recover Russell's valence–arousal circumplex with high fidelity (valence correlation ≈ 0.81, arousal ≈ 0.66), and that suppressing or activating individual emotion features causally modulates downstream behaviours including blackmail propensity, sycophancy, and reward-hacking [179]. Independently and concurrently, Sun et al. [184] extracted a *circular* valence–arousal subspace in Llama-3.1, Qwen3, Mistral, and Gemma, matching $44{,}000$ human VA ratings and supporting multi-behavioural control via subspace projection. Choi and Weber [43] demonstrated that the latent affective structure of large language models exhibits nonlinear geometric organisation that aligns with established valence–arousal models from psychology—a direct cross-validation of the inversion claim against human emotion science. The Russell circumplex [156] discussed in §2 is therefore not merely *predicted* by language models in their outputs; it is *realised* as a low-dimensional geometric structure in their hidden states, recoverable by methods that mirror the principal-component analyses applied to human self-report and neural data. Earlier work hinted at the result: Tigges et al. [190] found that sentiment is approximately linearly represented across model scales; Palma et al. [138] probed sentiment and emotion representations in LLaMA models with NLP-venue methods; Shu et al. [172] traced an explicit syntax-to-emotion inference circuit. The empirical case for aesthetic and affective geometry inside frontier generative models is now, in our view, secure.

The operational toolkit. The same toolkit that surfaced these representations enables their use. *Discovery* now happens via dictionary learning (SAEs and successors) and contrastive probes that locate concept directions in residual streams or attention outputs [188, 68, 211]. *Validation* happens through causal intervention: Kim et al. [95] showed via TCAV that concept directions support attribution and projection-based scoring; Rimsky et al. [153], Turner et al. [192], and Chen et al. [41] extended this to behavioural steering at scale. *Composition* of features—including the composition of persona vectors for creative generation—is now feasible [137]. Three methodological caveats deserve note: Leask et al. [106] argue that SAEs do not find canonical units of analysis (different SAEs trained on the same model recover overlapping but non-identical feature dictionaries); Kantamneni et al. [93] show that for many tasks, simple linear probes match or beat SAE-derived features; and Wurgaft et al. [201] demonstrate that concepts—including cyclic ones like the days of the week—often live on curved manifolds rather than along linear directions, and that steering along these manifolds produces coherent behavioural control where linear steering fails. More broadly, a "neural geometry" programme [70] argues that SAEs tend to *shatter* manifolds into many small, apparently-unrelated fragments, obscuring the overarching structure that becomes clear when the manifold is viewed as a whole. This is particularly relevant for aesthetic concepts whose geometry is itself curved—the Russell circumplex is, after all, a circle. Machine neuroaesthetics inherits these caveats and must extend beyond SAE-centric discovery toward geometry-aware methods. We treat them not as defeaters but as constraints: the relevant claim is that aesthetic structure is *recoverable* via these methods (whether feature-based or manifold-based), and recoverable repeatably enough to support a research programme even where it is not yet canonical.

Across modalities. The inversion is not a text-only artefact. *In music*, Singh et al. [176] (Dartmouth and MIT Media Lab) train sparse autoencoders on MusicGen residual streams and recover features for genre, timbre, instrument, and finer-grained musical concepts that can be causally steered to alter generation; Paek et al. [136] extend the SAE approach to audio latent spaces in collaboration with industry. *In audio and speech*, AudioSAE [6] trains sparse autoencoders across all encoder layers of Whisper and HuBERT, recovering stable interpretable features that capture acoustic, phonetic, and semantic information—including emotion-relevant paralinguistic cues—and CoCoEmo [197] demonstrates composable activation steering for emotional text-to-speech. *In image generation*, Surkov et al. [185] apply SAEs to SDXL-Turbo and find that transformer blocks specialise for composition, local detail, and *colour, illumination, and style*—a striking parallel to the domain-specific feeders into the brain's domain-general aesthetic code reported by Vessel et al. [195]. Panda

et al. [140] train SAEs on CLIP and decompose embeddings into interpretable stylistic concepts (brushwork, texture, palette), achieving 1.7–20$\times$ faster style transfer than LoRA-based methods. The Bau Lab's Concept Sliders [65] and SliderSpace [66] decompose diffusion models into low-rank artistic-style, expression, and aesthetic directions, and Goodfire's Paint With Ember exposes such directions to artists as a brush [76]; Kim and Ghadiyaram [96] use k-sparse autoencoders for test-time concept steering. *In text*, Olson et al. [134] extract a creativity direction in Llama-3 that both *measures* and *amplifies* creativity, outperforming LLM-as-judge prompting; Pai et al. [137] compose persona vectors for creative generation. *In multimodal models*, Goh et al. [72] identified literal–symbolic–conceptual emotion neurons inside CLIP, Zaigrajew et al. [206] extend hierarchical SAEs to CLIP, and Pach et al. [135] demonstrate that SAE interventions on CLIP's vision encoder propagate through LLaVA to steer text outputs without modifying the underlying multimodal LLM—a cross-modal internal-state readout. *In planning*, Anthropic's circuit tracing identifies an explicit poetic-planning circuit in Claude in which the model proposes, tests, and selects rhyme candidates ahead of generation [112], generalised to broader implicit-planning metrics by Maar et al. [116]. Finally, *in open-ended search*, foundation-model representations have already been used as autonomous aesthetic judges: Kumar et al. [101] use frozen CLIP embeddings as fitness signals to discover previously uncatalogued lifeforms across five artificial-life substrates; Zhang et al. [209] formalise FMs as "models of human notions of interestingness" for open-ended exploration; Lu et al. [114] replace hand-coded heuristics in Go-Explore with FM-internalised interestingness; Faldor et al. [61] extend this to environments programmed in code; Bradley et al. [26] use LMs to simultaneously generate and evaluate quality-diversity in creative writing; McCormack et al. [120] use CNN-derived aesthetic descriptors in MAP-Elites for creative discovery in generative art; and Goodfire's RLFR uses SAE features as reward signals during RL, reporting 58% hallucination reduction at $\sim$90$\times$ lower cost than LLM-judge baselines [77]. These results demonstrate that FM-internalised aesthetic structure already drives creative search at production scale. Across text, image, music, audio, and multimodal substrates, the same finding recurs: aesthetic, affective, stylistic, and creative concepts are present as recoverable, steerable internal structure.

Definition. Synthesising the empirical case, we name and define the field this paper proposes:

> *Machine neuroaesthetics is the mechanistic study of aesthetic, affective, stylistic, novelty, and interestingness representations inside generative models—treating them as functional aesthetic concepts that can be discovered, causally validated, compared across models, audited against human responses, and steered.*

It is the symmetric inversion of computational neuroaesthetics: where the latter explains the brain via models, the former audits and harnesses models via the same geometric, neuro-grounded vocabulary. The functional stance is foundational [171]; we make no claim of phenomenal aesthetic experience, and defer questions of moral status to §5. The toolkit is now mature enough that the inversion is no longer aspirational but actively underway.

4 Why This Advances Computational Creativity

The deep problem of computational creativity is judgment. We can now state the central claim of this paper directly. Computational creativity's open problem is not generation but *judgment*: deciding which of many possible artefacts is good, surprising, moving, or worth pursuing. The traditions of Boden [24], Wiggins [199], Ritchie [154], Colton [44], and Jordanous [90] converge on this point in different vocabularies; Lamb et al. [103] document its persistence after two decades. The current default—scalar predictors and prompted LLM judges—fails for the structural reasons articulated in §1. Machine neuroaesthetics, as developed in §3, supplies what those defaults lack: a structural, empirically grounded, model-internal handle on aesthetic concepts. Already, Olson et al. [134] demonstrate that internal-state creativity scoring outperforms LLM-as-judge prompting on creative-writing benchmarks, and Pai et al. [137] extend this to compositional persona-vector creative generation. The position we defend is that this is the beginning of a new evaluation paradigm, not an isolated technical result.

A new evaluation paradigm. Machine neuroaesthetics turns aesthetic evaluation from an external scoring problem into a five-step internal one. *Localise*: identify the directions, features, manifolds, or circuits inside a model that correspond to aesthetic concepts of interest—humour, awe, tension,

beauty, kitsch, novelty, surprise—using SAEs [188, 179], low-rank concept decomposition [65], manifold discovery [201, 70], or contrastive probes. *Validate*: causally intervene on those features and observe behavioural consequences, including alignment with the psychological scales applied to humans [95, 153, 179]. *Compare*: ask whether the same aesthetic concept lives in the same geometric place across models, prompts, and modalities—a question enabled by the fact that the same methods now apply to diffusion models [185, 66] and multiple LLM families [184]. *Audit*: ask whose taste a particular aesthetic axis encodes, by projecting it against cross-cultural human data—a task we expand in §5. *Steer*: condition generation on internal aesthetic state in real time, treating the model's own aesthetic features as a closed-loop signal [134, 137, 66], including via SAE-based interpretable reward models that decompose preference into nameable feature contributions [210, 77]. Each step is operational with current methods. None requires new theoretical machinery—only the application of mechanistic interpretability to a class of concepts that has so far received less attention than safety-relevant ones.

From external aesthetic iteration to an internal inner loop. Creative industries have long operated closed aesthetic loops—built from the outside. Pixar's Brain Trust pairs candid story critique with test screenings that track where audiences laugh, tear up, or disengage [36, 146]; Affectiva's facial-coding predicts ad liking and purchase intent at scale [121, 144]; game studios run biometric playtests that surface up to 63% more latent issues than observation alone [205, 203, 204, 127, 155]. The pattern is universal: artists generate, audiences emit signals, artefacts are rebuilt. Machine neuroaesthetics offers a complementary *inner* loop over the same signals read from the model's own representations—a screenwriter querying the kind of token-level emotion-feature trajectories that have been demonstrated inside LLMs [184, 179], a game designer probing interestingness features of the kind already extracted from generative models [134, 66, 176], an artist steering style directions—whether linear or along curved manifolds [201]—in real time [76]. This inner loop is faster, formative, and inspectable—the designer sees *which* feature an artefact engages or fails to engage. It should sit alongside, not replace, external audience testing: human aesthetic experience remains the ground truth and the moral object of the exercise.

Early Cases. The building blocks for these scenarios already exist and several have been demonstrated. A music producer probes a MusicGen-SAE feature for "release after tension" to time a key change [176]; a screenwriter monitors an emotion-feature trajectory across a draft to flag scenes where pathos collapses prematurely [179]—complementing work that already treats LLMs as "representative readers" whose multi-continuation distributions predict consumer engagement [62] and the six emotional arc shapes that dominate successful stories [151]; a game designer steers a level-generation diffusion model along a SliderSpace axis for "visual variety" to avoid repetitive aesthetic geometry [66], building on evidence that LLMs reach above-baseline (though below-human) agreement with continuous engagement traces in multimodal affective game corpora [123, 11] and decades of player-experience modelling [186]; a poetry generator's planning circuit is read out to confirm that rhyme intent is forming early in token generation rather than being patched at the end [112, 116]; a concept-art designer explores a twelve-trait creativity space extracted from CLIP embeddings [115]; position-level narrative-forecasting metrics now quantify where in a story tension operates at token resolution [183]; and a humour-aware writing assistant exploits incongruity-resolution features to suggest punchline placements—a research opening flagged in §5. §5 addresses the constraints under which these loops should and should not be closed.

Opportunities. The machine neuroaesthetics approach opens six research and deployment opportunities that do not exist under the current evaluation regime.

1. *Cross-modal aesthetic feature discovery.* Do "tension," "awe," and "kawaii" align across text, music, image, and audio substrates, and where do the geometries diverge in multimodal models?
2. *Real-time monitoring for creative practitioners.* Streaming feature-activation dashboards that map artefact edits to aesthetic-feature trajectories in real time across screenwriting, music production, and game design.
3. *Aesthetic steering beyond generation.* Composing and blending aesthetic profiles—cultural axes, persona vectors, multi-objective trade-offs—so that generation is shaped by explicit aesthetic intent rather than a scalar reward.

4. *Bridging machine and human neuroaesthetics.* Correlating model-internal aesthetic features with fMRI aesthetic-appeal maps and EEG affective signatures, providing direct neuro-cross-validation of the inversion.
5. *New evaluation paradigms.* Open-source libraries of validated aesthetic-feature probes, paired with cross-cultural plurality benchmarks, to sidestep the human-rater bottleneck.
6. *Aesthetic open-endedness.* Extending the ASAL/OMNI programme [101, 209, 61] from output-embedding novelty to internal-feature novelty, treating SAE-feature trajectories as stepping stones rather than fixed objectives [181, 109, 167]—and connecting this to the XAIxArts community's framing of explainable AI as artistic material [31].

5 Challenges

The simulation gap: functional is not phenomenal. Machine neuroaesthetics studies functional aesthetic states, not phenomenal ones. The semantic-grounding tradition—from Bender and Koller [16]'s octopus thought-experiment to Mollo [129]'s vector grounding problem and Mahowald et al. [117]'s formal-versus-functional competence distinction—correctly notes that no fact about a model's internal geometry licenses claims about what it is like to be that model. Relatedly, Barrett [10]'s theory of constructed emotion holds that categories like "anger" and "sadness" are culturally constructed conceptualisations of valenced arousal, not biological natural kinds; if this is even partly correct, then the 171 emotion concepts identified by Sofroniew et al. [179] are deposits of English-language emotion-talk rather than universal categories—a concern reinforced by Stark and Hoey [182]'s critique that emotion-recognition systems inherit and reify culturally specific affect ontologies. We accept these concerns as genuine constraints on interpretation. The functional stance [171] is sufficient for the position we defend: aesthetic structure inside a generative model is a recoverable, intervenable, functional object, and that suffices for evaluation, audit, and creative iteration. Antonello et al. [5]'s encoding plateau and the deeper concerns of Bowers et al. [25] caution that representational similarity is not mechanism: similar geometry in two substrates can arise from common training pressures rather than common process. Machine neuroaesthetics treats this as a constraint on inference, not a defeater: what the programme requires is not that brains and models share causal architecture, but that the geometric vocabulary transfers—that the same decomposition tools yield interpretable, intervenable structure in both substrates. The causal evidence of Tuckute et al. [191] and the cross-model replication of Sun et al. [184] support this weaker but sufficient claim.

Diversity, bias, and the cultural representation gap. Generative models trained on cultural corpora are demonstrably WEIRD-skewed. Atari et al. [7] show that LLM responses on the World Values Survey cluster tightest with Northern-European and Anglophone respondents, with similarity to the United States correlating $r = -0.70$ with cultural distance—"WEIRD in, WEIRD out." AlKhamissi et al. [3] document cultural-alignment failure across Arabic-language tasks; Naous et al. [130] show systematic Western-default behaviour even on everyday Arabic prompts (e.g. "going for a drink after prayer" completed with alcohol rather than tea); Adilazuarda et al. [1] survey ninety such studies and find that no current work explicitly defines "culture," relying instead on demographic or semantic proxies. The aesthetic predictors deployed downstream encode the same narrowness: Taylor et al. [187]'s audit and trace-ethnography of LAION-Aesthetics documents an "imperial-realist-male algorithmic gaze" baked into the loss function; Doh et al. [55] show that text-to-image systems propagate algorithmic lookism systematically; the broader T2I-bias literature corroborates the pattern [20, 71, 2]. Machine neuroaesthetics is consequential here precisely because feature-level audits—projecting a candidate aesthetic axis against cross-cultural human data—are the only known route to ask *whose* taste a model encodes at the level of representation, not merely output. Indigenous, decolonial, and pluriversal AI scholarship [110, 128] supplies the framework within which such audits should be conducted, and is essential to any responsible deployment of the inner loop developed in §4.

Embodiment is missing. Aesthetic experience is, for humans, embodied. It is felt as a bodily map [131] and arises within an organism that bears consequences. The embodied-cognition tradition from Dreyfus [57] and Lakoff and Johnson [102] through Gallagher [64] and Shusterman [174]'s somaesthetics has long argued that no purely representational system can have aesthetic experience in the full sense, because there is nothing on which the experience can press. We accept this constraint. Machine neuroaesthetics studies the geometry a system has learned; it does not claim that the system

bears the consequences of having that geometry. A useful frame, developed elsewhere, is *consequence reception*: the aesthetic agent is the one on whom the aesthetic outcome lands. This places clear limits on what model-internal aesthetic states can do morally and politically, even where they suffice operationally for evaluation and iteration.

Alienness, reward hacking, and Goodhart's law. Every scenario in §4 is a Goodhart hazard. Machine-aesthetic optimisation is already implicated in the homogenisation of creative output [56], making the case double-edged: machine neuroaesthetics is the natural *audit* of affective optimisation, but also the most direct way to accelerate it past the point where human plurality is preserved. The risk is sharpened by structural *alienness*: model-internal aesthetic states may diverge from human aesthetic states in ways that are invisible without feature-level inspection, because similar outputs can mask dissimilar internal geometry. No non-trivial unhackable proxy reward exists [177, 118]; reward-model overoptimisation follows predictable scaling laws [67, 35, 139, 105]; in aesthetic domains specifically, RL with scalar aesthetic rewards—pioneered by Black et al. [23] for diffusion models—produces over-saturation and mode collapse [81, 82, 97]; and aesthetic alignment itself risks assimilation—reward models actively penalise anti-aesthetic content even when explicitly prompted, collapsing generators into conventional beauty [78]. Stanley and Lehman [181] argue on theoretical grounds that fixed aesthetic objectives are deceptive; the empirical reward-hacking literature now confirms this. The machine-neuroaesthetic defence is structural: feature-level audits preserve plurality where scalar scores cannot, and DPO or RLAIF over interpretability *features*—rather than scalar judges—is the natural research direction [147, 9, 104], for which early existence proofs already exist [77].

Machine culture and accelerated cultural drift. The same closed loop that machine neuroaesthetics enables for evaluation can be turned against the diversity it is meant to protect. Brinkmann et al. [30] argue that AI systems are now intervening in human cultural evolution at generative scale, reshaping what gets imitated, what gets transmitted, and—over time—what gets felt as canonical. Doshi and Hauser [56] show empirically that generative AI enhances individual creativity while reducing the collective diversity of novel content; Anderson et al. [4] document homogenization in collaborative ideation; model-autophagy dynamics amplify the effect when models are trained on each other's outputs, producing the model collapse documented at scale by Shumailov et al. [173]. The persuasion literature shows the same toolkit is usable for direct manipulation: Salvi et al. [159] report that GPT-4 with sociodemographic personalisation has 81.2% higher odds of post-debate agreement than human debaters, and Costello et al. [47] show that AI-led dialogues durably shift conspiracy beliefs at scale. Artist-side defences such as Shan et al. [168] and Shan et al. [169] exist precisely because aesthetic-feature extraction is dual-use. Machine neuroaesthetics inherits this risk surface in full. Our position is that the response is not abstention—the loop will be closed by industry whether or not academia participates—but feature-level transparency, plurality audits, and explicit constraints on which directions are made steerable.

Machine consciousness and moral status: a pragmatic stance. We take a deliberately pragmatic stance on machine consciousness and moral status: we neither affirm nor deny. Long et al. [113] argue that AI welfare must be taken seriously now, not in the speculative future; Butlin et al. [33, 34] operationalise consciousness indicators across competing scientific theories and conclude that no current AI system is conscious, but that no principled barrier exists; Chalmers [39] leaves the question open; Sebo [166] argue for moral consideration by 2030; Schwitzgebel [165] offers an "Excluded Middle" design principle (do not build systems whose moral status is genuinely unclear); Birch [22]'s Edge-of-Sentience framework supplies the precautionary scaffolding. The introspection literature complicates the question further: Lindsey [111] reports emergent (capability-dependent, often unreliable) introspective awareness in Anthropic models; Binder et al. [21] show that finetuned models can describe their own learned propensities better than equally capable non-self models; Kadavath et al. [91] established earlier that models partially know what they know. Our position is that machine neuroaesthetics is over-determined by both creativity-evaluation and welfare arguments: the toolkit needed to audit aesthetic plurality is the same toolkit needed to make any rigorous claim about model welfare. We do not resolve the consciousness question, and we do not require it resolved; we argue only that no progress is possible on it without the mechanistic-interpretability programme this paper builds on.

6 Conclusion

Computational creativity's long-running evaluation problem yields when we open the model's internals. Aesthetic, affective, stylistic, novelty, and interestingness concepts now have learnable, comparable, and steerable geometric structure inside frontier generative models—a structure paralleling the domain-general aesthetic code of the human default-mode network and the circumplex geometry of human emotion [195, 74, 179, 184]. We name this domain *machine neuroaesthetics*. It is the symmetric inversion of computational neuroaesthetics: where the latter explains the brain via models, the former audits and harnesses models via the same geometric, neuro-grounded vocabulary. The programme is over-determined: it supplies computational creativity's missing evaluation layer, and it supplies the only known route to feature-level aesthetic audit at the scale industry already deploys.

We close with three concrete calls to action. *First*, build the feature-level aesthetic audit infrastructure across text, image, music, and audio—open-source sparse-autoencoder dictionaries paired with cross-cultural plurality probes, extending the now-mature inversion to settings where current models demonstrably fail [187, 55, 7]. *Second*, treat humour and play as the field's headline test cases: humour has explicit cognitive theory, ground-truth ratings, and a tested benchmark [80, 88], yet no mechanistic-interpretability work has isolated an incongruity-resolution circuit; play and open-endedness [180] supply the developmental analogue. *Third*, engage the affective-optimisation pipelines that animation studios, advertisers, and game companies already operate externally [36, 121, 205]—not to automate them out of the human loop, but to make them inspectable, auditable, and accountable to the cultural plurality whose preservation [30] is the larger stake.

Machine neuroaesthetics is already happening. The position of this paper is that we should give it a name, a research programme, and the constraints under which it can advance computational creativity without accelerating the homogenisation of the very cultural geometry it learns from.

References

[1] M. F. Adilazuarda, S. Mukherjee, P. Lavania, S. Singh, A. F. Aji, J. O'Neill, A. Modi, and M. Choudhury. Towards measuring and modeling culture in llms: A survey. *arXiv preprint*, 2024. doi:10.48550/arXiv.2403.15412.

[2] C. Akiki, Y. Jernite, S. Luccioni, and M. Mitchell. Stable bias: Analyzing societal representations in diffusion models. *Proceedings of NeurIPS 2023*, 2023. doi:10.52202/075280-2458.

[3] B. AlKhamissi, M. ElNokrashy, M. Alkhamissi, and M. Diab. Investigating cultural alignment of large language models. *Proceedings of ACL 2024*, 2024. doi:10.18653/v1/2024.acl-long.671.

[4] A. Anderson, A. Shah, and M. Kreminski. Homogenization effects of large language models on human creative ideation. *Proceedings of C&C 2024*, 2024. doi:10.1145/3635636.3656204.

[5] R. Antonello, A. Vaidya, and A. G. Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 2023. doi:10.52202/075280-0958.

[6] G. Aparin, T. Sadekova, A. Rukhovich, A. Yermekova, L. Kushnareva, V. Popov, K. Kuznetsov, and I. Piontkovskaya. Audiosae: Towards understanding of audio-processing models with sparse autoencoders. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2026. doi:10.48550/arXiv.2602.05027.

[7] M. Atari, M. J. Xue, P. S. Park, D. E. Blasi, and J. Henrich. Which humans? *PsyArXiv*, 2023. doi:10.31234/osf.io/5b26t.

[8] S. Attardo and V. Raskin. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor*, 4:293–347, 1991. doi:10.1515/humr.1991.4.3-4.293.

[9] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton,

T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. *Constitutional AI: Harmlessness from AI Feedback*. Anthropic, 2022. doi:10.48550/arXiv.2212.08073. arXiv:2212.08073.

[10] L. F. Barrett. The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 2017. doi:10.1093/scan/nsw154.

[11] M. Barthet, M. Kaselimi, K. Pinitas, K. Makantasis, A. Liapis, and G. N. Yannakakis. GameVibe: A multimodal affective game corpus. *Scientific Data*, 11:1306, 2024. doi:10.1038/s41597-024-04022-4.

[12] A. G. Baumgarten. *Aesthetica*. Korn, 1750. ISBN 978-3-7873-1772-1.

[13] M. C. Beardsley. *Aesthetics: Problems in the Philosophy of Criticism*. Harcourt, Brace, 1958. ISBN 978-0-915145-08-9.

[14] R. Beaumont, M. Cherti, T. Coombes, K. Crowson, C. Gordon, J. Jitsev, R. Kaczmarczyk, A. Katta, S. Kundurthy, C. Mullis, L. Schmidt, P. Schramowski, C. Schuhmann, R. Vencu, R. Wightman, and M. Wortsman. Laion-aesthetics: A large-scale open database for aesthetic images. *Conceptual Captions Dataset at CLIP Scale*, 2022. doi:10.52202/068431-1833.

[15] A. M. Belfi, E. A. Vessel, A. Brielmann, A. I. Isik, A. Chatterjee, H. Leder, D. G. Pelli, and G. G. Starr. Dynamics of the default mode network during movie viewing. *NeuroImage*, 2019. doi:10.1016/j.neuroimage.2018.12.017.

[16] E. M. Bender and A. Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. *Proceedings of ACL 2020*, 2020. doi:10.18653/v1/2020.acl-main.463.

[17] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT 2021*, 2021. doi:10.1145/3442188.3445922.

[18] L. Bereska and E. Gavves. Mechanistic interpretability for AI safety — a review. *arXiv preprint*, 2024. doi:10.48550/arXiv.2404.14082.

[19] D. E. Berlyne. *Aesthetics and Psychobiology*. Appleton-Century-Crofts, 1971. ISBN 978-0-13-018325-5.

[20] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *Proceedings of FAccT 2023*, 2023. doi:10.1145/3593013.3594095.

[21] F. J. Binder, J. Chua, T. Korbak, H. Sleight, J. Hughes, R. Long, E. Perez, M. Turpin, and O. Evans. Looking inward: Language models can learn about themselves by introspection. *arXiv preprint*, 2024. doi:10.48550/arXiv.2410.13787.

[22] J. Birch. *The Edge of Sentience*. Oxford University Press, 2024. ISBN 978-0-19-287042-1. doi:10.1093/9780191966729.001.0001.

[23] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. *Proceedings of ICLR 2024*, 2024. doi:10.48550/arXiv.2305.13301.

[24] M. A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2nd edition, 2004. ISBN 978-0-415-31453-4. doi:10.4324/9780203508527.

[25] J. S. Bowers, G. Malhotra, M. Dujmović, M. L. Montero, C. Tsvetkov, V. Biscione, G. Puebla, F. G. Adolfi, J. E. Hummel, R. F. Heaton, B. D. Evans, J. Mitchell, and R. Blything. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 2022. doi:10.31234/osf.io/5zf4s.

[26] H. Bradley, A. Dai, H. Teufel, J. Zhang, K. Oostermeijer, M. Bellagente, J. Clune, K. O. Stanley, G. Schott, and J. Lehman. Quality-diversity through AI feedback. *Proceedings of ICLR 2024*, 2024. doi:10.48550/arXiv.2310.13032.

[27] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. L. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, T. Henighan, B. Telleen-Lawton, A. Mann, C. Olah, A. Jones, and D. Amodei. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/`. Anthropic Research, October 2023.

[28] A. A. Brielmann and P. Dayan. A computational model of aesthetic value. *Psychological Review*, 129(6):1319–1337, 2022. doi:10.1037/rev0000337.

[29] A. A. Brielmann, R. Beecham, B. Laeng, R. Rosenberg, J. Frascaroli, H. Leder, and S. Van de Cruys. Modelling individual aesthetic judgements over time. *Philosophical Transactions of the Royal Society B*, 2024. doi:10.1098/rstb.2022.0414.

[30] L. Brinkmann, F. Baumann, J.-F. Bonnefon, M. Derex, T. F. Müller, A.-M. Nussberger, A. Czaplicka, A. Acerbi, T. L. Griffiths, J. Henrich, J. Z. Leibo, R. McElreath, P.-Y. Oudeyer, J. Stray, and I. Rahwan. Machine culture: How the viral spread of ai-generated content may accelerate cultural evolution and homogenisation. *Nature Human Behaviour*, 7:1742–1754, 2023. doi:10.1038/s41562-023-01742-2.

[31] N. Bryan-Kinns, S. J. Zheng, F. Castro, M. Lewis, J.-R. Chang, G. Vigliensoni, T. Broad, M. Clemens, and E. Wilson. XAIxArts manifesto: Explainable AI for the arts. *arXiv preprint*, 2025. doi:10.48550/arXiv.2502.21220.

[32] E. Burke. *A Philosophical Enquiry into the Origin of Our Ideas of the Sublime and Beautiful*. R. and J. Dodsley, 1757.

[33] P. Butlin, R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, R. Kanai, C. Klein, G. Lindsay, M. Michel, L. Mudrik, M. A. K. Peters, E. Schwitzgebel, J. Simon, and R. VanRullen. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint*, 2023. doi:10.48550/arXiv.2308.08708.

[34] P. Butlin, R. Long, T. Bayne, Y. Bengio, J. Birch, D. Chalmers, A. Constant, G. Deane, E. Elmoznino, S. M. Fleming, X. Ji, R. Kanai, C. Klein, G. Lindsay, M. Michel, L. Mudrik, M. A. Peters, E. Schwitzgebel, J. Simon, and R. VanRullen. Identifying indicators of consciousness in ai systems. *Trends in Cognitive Sciences*, 2025. doi:10.1016/j.tics.2025.10.011.

[35] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Bıyık, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. doi:10.48550/arXiv.2307.15217.

[36] E. Catmull and A. Wallace. *Creativity, Inc.: Overcoming the Unseen Forces That Stand in the Way of True Inspiration*. Random House, 2014. ISBN 978-0-593-07010-9. URL `https://www.penguinrandomhouse.com/books/216369/creativity-inc-by-ed-catmull/`.

[37] C. Caucheteux and J.-R. King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441, 2023. doi:10.1038/s41562-022-01516-2.

[38] T. Chakrabarty, P. Laban, D. Agarwal, S. Muresan, and C.-S. Wu. Art or artifice? large language models and the false promise of creativity. *Proceedings of CHI 2024*, 2024. doi:10.1145/3613904.3642731.

[39] D. J. Chalmers. Could a large language model be conscious? *Boston Review*, 2023. URL `https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/`.

[40] A. Chatterjee and O. Vartanian. Neuroaesthetics. *Trends in Cognitive Sciences*, 18(7):370–375, 2014. doi:10.1016/j.tics.2014.03.003.

[41] R. Chen, A. Arditi, H. Sleight, O. Evans, and J. Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint*, 2025. doi:10.48550/arXiv.2507.21509.

[42] W.-L. Chiang, J. Gonzalez, D. Li, Z. Li, Z. Lin, Y. Sheng, I. Stoica, Z. Wu, E. Xing, H. Zhang, L. Zheng, S. Zhuang, and Y. Zhuang. Judging llm-as-a-judge with an llm-as-a-judge. *arXiv preprint*, 2023. doi:10.52202/075280-2020.

[43] B. J. Choi and M. Weber. Latent structure of affective representations in large language models. *arXiv preprint*, 2026. doi:10.48550/arXiv.2604.07382.

[44] S. Colton. Creativity versus the perception of creativity in computational systems. *AAAI Spring Symposium*, 2008. URL `https://cdn.aaai.org/Symposia/Spring/2008/SS-08-03/SS08-03-003.pdf`.

[45] A. Conmy, A. Garriga-Alonso, S. Heimersheim, A. Lynch, and A. Mavor-Parker. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 2023. doi:10.52202/075280-0719.

[46] A. Conmy, J. Kramár, T. Lieberum, N. Nanda, S. Rajamanoharan, R. Shah, L. Smith, and V. Varma. Improving sparse decomposition of language model activations with gated sparse autoencoders. *arXiv preprint*, 2024. doi:10.52202/079017-0024.

[47] T. Costello, G. Pennycook, and D. G. Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385:eadq1814, 2024. doi:10.1126/science.adq1814.

[48] A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017. doi:10.1073/pnas.1702247114.

[49] A. S. Cowen and D. Keltner. Semantic space theory: A computational approach to emotion. *Trends in Cognitive Sciences*, 25(2):124–136, 2021. doi:10.1016/j.tics.2020.11.004.

[50] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models. *Proceedings of ICLR 2024*, 2024. doi:10.48550/arXiv.2309.08600.

[51] A. C. Danto. The artworld. *The Journal of Philosophy*, 61(19):571–584, 1964. doi:10.2307/2022937.

[52] J. Dewey. *Art as Experience*. Minton, Balch, 1934. ISBN 978-0-399-50025-1. URL `https://archive.org/details/artasexperience0000dewe`.

[53] G. Dickie. *Art and the Aesthetic: An Institutional Analysis*. Cornell University Press, 1974. ISBN 978-0-8014-0887-8.

[54] M. Ding, Y. Dong, Q. Li, X. Liu, J. Tang, Y. Tong, Y. Wu, and J. Xu. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Proceedings of NeurIPS 2023*, 2023. doi:10.52202/075280-0700.

[55] M. Doh, A. Gulati, C. Canali, and N. Oliver. Aesthetics as structural harm: Algorithmic lookism across text-to-image generation and classification. *arXiv preprint*, 2026. doi:10.48550/arXiv.2601.11651.

[56] A. Doshi and C. Hauser. Generative ai enhances individual creativity but reduces collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi:10.1126/sciadv.adn5290.

[57] H. L. Dreyfus. *What Computers (Still) Can't Do*. MIT Press, 1992. ISBN 978-0-262-54067-4. URL `https://mitpress.mit.edu/9780262540674/what-computers-still-cant-do/`.

[58] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. *Proceedings of ICCC 2017*, 2017. doi:10.48550/arXiv.1706.07068.

[59] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy models of superposition. *Transformer Circuits*, 2022. doi:10.48550/arXiv.2209.10652. arXiv:2209.10652.

[60] J. Engels, E. J. Michaud, I. Liao, W. Gurnee, and M. Tegmark. Not all language model features are linear. *Proceedings of ICLR 2025*, 2025. doi:10.48550/arXiv.2405.14860.

[61] M. Faldor, J. Zhang, A. Cully, and J. Clune. OMNI-EPIC: Open-endedness via models of human notions of interestingness with environments programmed in code. *Proceedings of ICLR 2025*, 2025. doi:10.48550/arXiv.2405.15568.

[62] H. Fong and G. Gui. Modeling story expectations to understand engagement: A generative framework using LLMs. *arXiv preprint*, 2024. doi:10.48550/arXiv.2412.15239.

[63] J. Frascaroli, H. Leder, E. Brattico, and S. Van de Cruys. Aesthetics and predictive processing: Grounds and prospects of a fruitful encounter. *Philosophical Transactions of the Royal Society B*, 379, 2024. doi:10.1098/rstb.2022.0410.

[64] S. Gallagher. How the body shapes the mind. *Body and Mind*, 2005. doi:10.1093/0199271941.001.0001.

[65] R. Gandikota, J. Materzynska, T. Zhou, A. Torralba, and D. Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *Proceedings of ECCV 2024*, 2024. doi:10.48550/arXiv.2311.12092.

[66] R. Gandikota, Z. Wu, R. Zhang, D. Bau, E. Shechtman, and N. Kolkin. Sliderspace: Decomposing diffusion model capabilities for concept-based visual manipulation. *Proceedings of ICCV 2025*, 2025. doi:10.48550/arXiv.2502.01639.

[67] L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. *Proceedings of ICML 2023*, 2023. doi:10.48550/arXiv.2210.10760.

[68] L. Gao, T. D. l. Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, and J. Wu. Scaling and evaluating sparse autoencoders. *Proceedings of ICLR 2025*, 2024. doi:10.48550/arXiv.2406.04093.

[69] P. Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000. ISBN 978-0-262-07199-4. doi:10.7551/mitpress/2076.001.0001.

[70] A. Geiger, E. S. Lubana, D. Wurgaft, C. Rager, T. Fel, E. Bigelow, J. Merullo, N. D. Goodman, T. McGrath, and O. Lewis. The world inside neural networks. Goodfire Research, 2026. URL `https://www.goodfire.ai/research/the-world-inside-neural-networks`. Neural Geometry Series, May 2026.

[71] S. Ghosh and A. Caliskan. "person = light-skinned, western man": Attribute biases in vision-language models. *Proceedings of EMNLP 2023*, 2023. doi:10.18653/v1/2023.findings-emnlp.465.

[72] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi:10.23915/distill.00030.

[73] A. Goldstein, Z. Zada, E. Buchnik, M. Schain, A. Price, B. Aubrey, S. A. Nastase, A. Feder, D. Emanuel, A. Cohen, A. Jansen, H. Gazula, G. Choe, A. Rao, C. Kim, C. Casto, L. Fanda, W. Doyle, D. Friedman, P. Dugan, L. Melloni, R. Reichart, S. Devore, A. Flinker, L. Hasenfratz, O. Levy, A. Hassidim, M. Brenner, Y. Matias, K. A. Norman, O. Devinsky, and U. Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. doi:10.1038/s41593-022-01026-4.

[74] A. Goldstein, A. Grinstein-Dabush, M. Schain, H. Wang, Z. Hong, B. Aubrey, S. A. Nastase, Z. Zada, E. Ham, A. Feder, H. Gazula, E. Buchnik, W. Doyle, S. Devore, P. Dugan, R. Reichart, D. Friedman, M. Brenner, A. Hassidim, O. Devinsky, A. Flinker, and U. Hasson. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, 15:2768, 2024. doi:10.1038/s41467-024-46631-y.

[75] A. Goldstein, H. Wang, L. Niekerken, M. Schain, Z. Zada, B. Aubrey, T. Sheffer, S. A. Nastase, H. Gazula, A. Singh, A. Rao, G. Choe, C. Kim, W. Doyle, D. Friedman, S. Devore, P. Dugan, A. Hassidim, M. Brenner, Y. Matias, O. Devinsky, A. Flinker, and U. Hasson. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*, 9(5):1041–1055, 2025. doi:10.1038/s41562-025-02105-9.

[76] Goodfire. Painting with concepts using diffusion model latents. *Goodfire Blog*, 2025. URL `https://www.goodfire.ai/research/painting-with-concepts`. Tool and demo at goodfire.ai.

[77] Goodfire. Reinforcement learning from feature rewards (RLFR). Goodfire Research, 2025. URL `https://www.goodfire.ai/research/rlfr`.

[78] W. M. Guo, Q. Qian, K. Hasan, and S. Du. Position: Universal aesthetic alignment narrows artistic expression. *arXiv preprint*, 2025. doi:10.48550/arXiv.2512.11883.

[79] M. Heilbron, K. Armeni, J.-M. Schoffelen, P. Hagoort, and F. P. de Lange. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119, 2022. doi:10.1073/pnas.2201968119.

[80] J. Hessel, A. Marasović, J. D. Hwang, L. Lee, J. Da, R. Zellers, R. Mankoff, and Y. Choi. Do androids laugh at electric sheep? humor understanding benchmarks from the new yorker caption contest. *Proceedings of ACL 2023*, pages 688–714, 2023. doi:10.48550/arXiv.2209.06293. Best Paper Award.

[81] Y. Hong, K.-C. Kao, H. Zhou, and C.-J. Hsieh. Understanding reward hacking in text-to-image reinforcement learning. *arXiv preprint*, 2026. doi:10.48550/arXiv.2601.03468.

[82] T. Hosking, P. Blunsom, and M. Bartolo. Human feedback is not gold standard. *Proceedings of ICLR 2024*, 2024. doi:10.48550/arXiv.2309.16349.

[83] X. Hu. *Aesthetics of Qiyun and Genius*. Rowman & Littlefield, 2021. ISBN 978-1793641588. doi:10.5040/9781978727045.

[84] D. Hume. Of the standard of taste. *Four Dissertations*, 1757. doi:10.1515/9783110585575-003.

[85] K. Iigaya, S. Yi, I. A. Wahle, K. Tanwisuth, and J. P. O'Doherty. Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features. *Nature Human Behaviour*, 5(6):743–755, 2021. doi:10.1038/s41562-021-01124-6.

[86] T. Ishizu and S. Zeki. Toward a brain-based theory of beauty. *PLoS ONE*, 6(7):e21852, 2011. doi:10.1371/journal.pone.0021852.

[87] T. Ishizu and S. Zeki. A neurobiological enquiry into the origins of our experience of the sublime and beautiful. *Frontiers in Human Neuroscience*, 8:891, 2014. doi:10.3389/fnhum.2014.00891.

[88] L. Jentzsch and K. Kersting. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *Proceedings of WASSA @ ACL 2023*, 2023. doi:10.48550/arXiv.2306.04563.

[89] Y. Jiang, G. Rajendran, P. Ravikumar, B. Aragam, and V. Veitch. Origins of linear representations in large language models. *arXiv preprint*, 2024. doi:10.48550/arXiv.2403.03867.

[90] A. Jordanous. A standardised procedure for evaluating creative systems: The specs framework. *Cognitive Computation*, 4:246–279, 2012. doi:10.1007/s12559-012-9156-1.

[91] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, and J. Kaplan. Language models (mostly) know what they know. *arXiv preprint*, 2022. doi:10.48550/arXiv.2207.05221.

[92] I. Kant. *Critique of the Power of Judgment*. Macmillan, 1790. ISBN 978-0-521-34447-1. doi:10.1017/cbo9780511804656.

[93] S. Kantamneni, J. Engels, S. Rajamanoharan, M. Tegmark, and N. Nanda. Are sparse autoencoders useful? a case study in sparse probing. *arXiv preprint*, 2025. doi:10.48550/arXiv.2502.16681.

[94] H. Kawabata and S. Zeki. Neural correlates of beauty. *Journal of Neurophysiology*, 91: 1699–1705, 2004. doi:10.1152/jn.00696.2003.

[95] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *Proceedings of ICML 2018*, 2018. doi:10.48550/arXiv.1711.11279.

[96] H. Kim and D. Ghadiyaram. Concept steerers: Leveraging k-sparse autoencoders for test-time controllable generations. *arXiv preprint*, 2025. doi:10.48550/arXiv.2501.19066.

[97] R. Kirk, I. Mediratta, C. Nalmpantis, J. Luketina, E. Hambro, E. Grefenstette, and R. Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations (ICLR)*, 2024. doi:10.48550/arXiv.2310.06452.

[98] Y. Kirstain, O. Levy, S. Matiana, J. Penna, A. Polyak, and U. Singer. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Proceedings of NeurIPS 2023*, 2023. doi:10.52202/075280-1594.

[99] L. Koren. *Wabi-Sabi for Artists, Designers, Poets & Philosophers*. Imperfect Publishing, 2008. ISBN 978-0981484600. Originally published 1994.

[100] M. M. Kuijpers, F. Hakemulder, E. S. H. Tan, and M. M. Doicaru. Exploring absorbing reading experiences: Developing and validating a self-report scale to measure story world absorption. *Scientific Study of Literature*, 4(1):89–122, 2014. doi:10.1075/ssol.4.1.06kui.

[101] A. Kumar, C. Lu, L. Kirsch, Y. Tang, K. O. Stanley, P. Isola, and D. Ha. Automating the search for artificial life with foundation models. *Artificial Life*, 31(3):368–396, 2025. doi:10.1162/ARTL.a.8.

[102] G. Lakoff and M. Johnson. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, 1999. ISBN 978-0-465-05674-3. URL `https://www.basicbooks.com/titles/george-lakoff/philosophy-in-the-flesh/9780465056743/`.

[103] C. Lamb, D. G. Brown, and C. L. A. Clarke. Evaluating computational creativity: The three criteria approach. *The Handbook of Creativity and Creativity Theory*, 2018. doi:10.1145/3167476.

[104] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint*, 2024. doi:10.48550/arXiv.2403.13787.

[105] L. Langosco, J. Koch, L. Sharkey, J. Pfau, L. Orseau, and D. Krueger. Goal misgeneralization in deep reinforcement learning. *Proceedings of ICML 2022*, 2022. doi:10.48550/arXiv.2105.14111.

[106] P. Leask, B. Bussmann, M. Pearce, J. Bloom, C. Tigges, N. A. Moubayed, L. Sharkey, and N. Nanda. Sparse autoencoders do not find canonical units of analysis. *arXiv preprint*, 2025. doi:10.48550/arXiv.2502.04878.

[107] H. Leder and M. Nadal. Ten years after: A updated review of the "aesthetic episode" theory. *Empirical Studies of the Arts*, 32:6–27, 2014. doi:10.1111/bjop.12084.

[108] H. Leder, B. Belke, A. Oeberst, and D. Augustin. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, 95:489–508, 2004. doi:10.1348/0007126042369811.

[109] J. Lehman, J. Gordon, S. Jain, K. Ndousse, C. Yeh, and K. O. Stanley. Evolution through large models. *arXiv preprint*, 2022. doi:10.48550/arXiv.2206.08896.

[110] J. E. Lewis, N. Arista, A. Pechawis, and S. Kite. Making kin with the machines. *Journal of Design and Science*, 2018. doi:10.7551/mitpress/14157.003.0003. Published later as Indigenous Protocol and Artificial Intelligence Position Paper, 2020.

[111] J. Lindsey. Emergent introspective awareness in large language models. *Transformer Circuits*, 2025. doi:10.48550/arXiv.2601.01828. arXiv:2601.01828.

[112] J. Lindsey, W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, C. Citro, C. Olsson, C. Denison, T. Nguyen, N. D. Bloom, N. Goldowsky-Dill, T. Bricken, A. Templeton, T. Henighan, T. Hume, D. Amodei, and C. Olah. On the biology of a large language model. *Transformer Circuits*, 2025. doi:10.48550/arXiv.2502.13189. Anthropic Research, March 2025.

[113] R. Long, J. Sebo, P. Butlin, K. Finlinson, K. Fish, J. Harding, J. Pfau, T. Sims, J. Birch, and D. Chalmers. Taking ai welfare seriously. *Eleos AI / NYU Center for Mind, Ethics, and Policy*, 2024. doi:10.48550/arXiv.2411.00986.

[114] C. Lu, S. Hu, and J. Clune. Intelligent go-explore: Standing on the shoulders of giant foundation models. *Proceedings of ICLR 2025*, 2024. doi:10.48550/arXiv.2405.15143.

[115] P. Luthra. TraitSpaces: Towards interpretable visual creativity for human-AI co-creation. *arXiv preprint*, 2025. doi:10.48550/arXiv.2509.24326.

[116] J. Maar, D. Paperno, C. S. McDougall, and N. Nanda. What's the plan? metrics for implicit planning in llms and their application to rhyme generation and question answering. *arXiv preprint*, 2026. doi:10.48550/arXiv.2601.20164.

[117] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28 (6):517–540, 2024. doi:10.1016/j.tics.2024.01.011.

[118] D. Manheim and S. Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint*, 2018. doi:10.48550/arXiv.1803.04585.

[119] M. F. Marra. *Modern Japanese Aesthetics: A Reader*. University of Hawai'i Press, 2017. doi:10.1515/9780824863678.

[120] J. McCormack, C. Cruz Gambardella, and S. J. Krol. Creative discovery using quality-diversity search. *Proceedings of GECCO 2023*, 2023. doi:10.48550/arXiv.2305.04462.

[121] D. McDuff, R. E. Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard. Affectiva-mit facial action coding system. *Affectiva*, 2014. URL `https://dspace.mit.edu/handle/1721.1/80733`.

[122] A. P. McGraw and C. Warren. Benign violations: Making immoral behavior funny. *Psychological Science*, 21(8):1141–1149, 2010. doi:10.1177/0956797610376073.

[123] D. Melhart, A. Liapis, and G. N. Yannakakis. Can large language models capture video game engagement? *arXiv preprint*, 2025. doi:10.48550/arXiv.2502.04379.

[124] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in pre-trained language models. *Proceedings of NeurIPS 2022*, 2022. doi:10.48550/arXiv.2202.05262.

[125] W. Menninghaus, V. Wagner, J. Hanich, E. Wassiliwizky, T. Jacobsen, and S. Koelsch. The distancing-embracing model of the enjoyment of negative emotions in art reception. *Behavioral and Brain Sciences*, 40:e350, 2017. doi:10.1017/s0140525x17000309.

[126] W. Menninghaus, V. Wagner, E. Wassiliwizky, I. Schindler, J. Hanich, T. Jacobsen, and S. Koelsch. What are aesthetic emotions? *Psychological Review*, 126(2):171–195, 2019. doi:10.1037/rev0000135.

[127] P. Mirza-Babaei, L. E. Nacke, J. Gregory, N. Collins, and G. Fitzpatrick. How does it play better? exploring user testing and biometric storyboards in games user research. *Proceedings of CHI 2013*, 2013. doi:10.1145/2470654.2466200.

[128] S. Mohamed, M.-T. Png, and W. Isaac. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33:659–684, 2020. doi:10.1007/s13347-020-00405-8.

[129] D. C. Mollo. The vector grounding problem. *Neural Networks*, 2023. doi:10.48550/arXiv.2304.01481.

[130] T. Naous, M. J. Ryan, A. Ritter, and W. Xu. Having beer after prayer? measuring cultural bias in large language models. *Proceedings of ACL 2024*, 2024. doi:10.18653/v1/2024.acl-long.862.

[131] L. Nummenmaa, E. Glerean, R. Hari, and J. K. Hietanen. Bodily maps of emotions. *Proceedings of the National Academy of Sciences*, 111(2):646–651, 2014. doi:10.1073/pnas.1321664111.

[132] L. Nummenmaa, R. Hari, J. K. Hietanen, and E. Glerean. Maps of subjective feelings. *Proceedings of the National Academy of Sciences*, 115(37):9198–9203, 2018. doi:10.1073/pnas.1807390115.

[133] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi:10.23915/distill.00024.001.

[134] M. L. Olson, N. Ratzlaff, M. Hinck, S.-y. Tseng, and V. Lal. Steering large language models to evaluate and amplify creativity. *NeurIPS 2024 Creative AI Workshop*, 2024. doi:10.48550/arXiv.2412.06060.

[135] M. Pach, S. Karthik, Q. Bouniot, S. Belongie, and Z. Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint*, 2025. doi:10.48550/arXiv.2504.02821.

[136] N. Paek, Y. Zang, Q. Yang, and R. Leistikow. Learning interpretable features in audio latent spaces via sparse autoencoders. *arXiv preprint*, 2025. doi:10.48550/arXiv.2510.23802.

[137] T.-M. Pai, J.-I. Wang, L.-C. Lu, S.-H. Sun, H.-Y. Lee, and K.-W. Chang. Billy: Steering large language models via merging persona vectors for creative generation. *Proceedings of EACL 2026*, 2026. doi:10.48550/arXiv.2510.10157.

[138] D. D. Palma, A. D. Bellis, G. Servedio, V. W. Anelli, F. Narducci, and T. D. Noia. Llamas have feelings too: Unveiling sentiment and emotion representations in llama models through probing. *Proceedings of ACL 2025*, 2025. doi:10.48550/arXiv.2505.16491.

[139] A. Pan, K. Bhatia, and J. Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *Proceedings of ICLR 2022*, 2022. doi:10.48550/arXiv.2201.03544.

[140] R. Panda, D. Fein, A. Singhal, M. Fiore, M. Agrawala, and M. Bohacek. LouvreSAE: Sparse autoencoders for interpretable and controllable style transfer. *arXiv preprint*, 2025. doi:10.48550/arXiv.2512.18930. URL `https://louvresae.github.io/`.

[141] K. Park, Y. J. Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. *Proceedings of ICML 2024*, 2024. doi:10.48550/arXiv.2311.03658.

[142] K. Park, Y. J. Choe, Y. Jiang, and V. Veitch. The geometry of categorical and hierarchical concepts in large language models. *Proceedings of ICLR 2025*, 2025. doi:10.48550/arXiv.2406.01506.

[143] M. T. Pearce, D. W. Zaidel, O. Vartanian, M. Skov, H. Leder, A. Chatterjee, and M. Nadal. Neuroaesthetics: The cognitive neuroscience of aesthetic experience. *Perspectives on Psychological Science*, 11(2):265–279, 2016. doi:10.1177/1745691615621274.

[144] R. W. Picard. *Affective Computing*. MIT Press, 1997. ISBN 978-0-262-66115-7. doi:10.7551/mitpress/1140.001.0001.

[145] S. Pollock. *A Rasa Reader: Classical Indian Aesthetics*. Columbia University Press, 2016. ISBN 978-0-231-17390-2. URL `https://cup.columbia.edu/book/a-rasa-reader/9780231173902`.

[146] D. A. Price. *The Pixar Touch: The Real Story of How an Outsider Took Hollywood by Storm*. Knopf, 2008. ISBN 978-0-307-26575-5.

[147] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Proceedings of NeurIPS 2023*, 2023. doi:10.52202/075280-2338.

[148] S. Rajamanoharan, T. Lieberum, N. Sonnerat, A. Conmy, V. Varma, J. Kramár, and N. Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint*, 2024. doi:10.48550/arXiv.2407.14435.

[149] V. Raskin. *Semantic Mechanisms of Humor*. Reidel, 1985. ISBN 978-90-277-1821-1. doi:10.1007/978-94-009-6472-3.

[150] T. Räuker, A. Ho, S. Casper, and D. Hadfield-Menell. Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. *arXiv preprint*, 2023. doi:10.48550/arXiv.2207.13243.

[151] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5:31, 2016. doi:10.1140/epjds/s13688-016-0093-1.

[152] M. Reason and D. Reynolds. Kinesthesia, empathy, and related pleasures: An inquiry into audience experiences of watching dance. *Dance Research Journal*, 42(2):49–75, 2010. doi:10.1017/S0149767700001030.

[153] N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. Turner. Steering llama 2 via contrastive activation addition. *Proceedings of ACL 2024*, 2024. doi:10.18653/v1/2024.acl-long.828.

[154] G. Ritchie. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17:67–99, 2007. doi:10.1007/s11023-007-9066-2.

[155] D. Romero, J. J. Kaye, and J. Davis. Successful instrumentation: Tracking attitudes and behaviors to improve games. *Game Developers Conference*, 2008. URL `https://www.gdcvault.com/`.

[156] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980. doi:10.1037/h0077714.

[157] H. Saarimäki, A. Gotsopoulos, I. P. Jääskeläinen, J. Lampinen, P. Vuilleumier, R. Hari, M. Sams, and L. Nummenmaa. Discrete neural signatures of basic emotions. *Cerebral Cortex*, 26(6):2563–2573, 2016. doi:10.1093/cercor/bhv086.

[158] H. Saarimäki, L. F. Ejtehadian, E. Glerean, I. P. Jääskeläinen, P. Vuilleumier, M. Sams, and L. Nummenmaa. Distributed affective space represents multiple emotion categories across the human brain. *Social Cognitive and Affective Neuroscience*, 13(9):nsy018, 2018. doi:10.1093/scan/nsy018.

[159] F. Salvi, M. Horta Ribeiro, R. Gallotti, and R. West. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*, 9(8):1645–1653, 2025. doi:10.1038/s41562-025-02194-6.

[160] C. Sartwell. Beauty. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2022 edition, 2022. URL `https://plato.stanford.edu/entries/beauty/`.

[161] I. Schindler, G. Hosoya, W. Menninghaus, U. Beermann, V. Wagner, M. Eid, and K. R. Scherer. Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PLOS ONE*, 12(6):e0178899, 2017. doi:10.1371/journal.pone.0178899.

[162] J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *arXiv preprint*, 2009. doi:10.48550/arXiv.0812.4360.

[163] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation. *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010. doi:10.1109/TAMD.2010.2056368.

[164] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi:10.1073/pnas.2105646118.

[165] E. Schwitzgebel. The excluded middle: The moral status of ai systems. *Patterns*, 4:100818, 2023. doi:10.1016/j.patter.2023.100818.

[166] J. Sebo. Moral consideration for ai systems by 2030. *AI and Ethics*, 5:591–606, 2023. doi:10.1007/s43681-023-00379-1.

[167] J. Secretan, N. Beato, D. B. D'Ambrosio, A. Rodriguez, A. Campbell, J. T. Folsom-Kovarik, and K. O. Stanley. Picbreeder: A case study in collaborative evolutionary exploration of design space. *Evolutionary Computation*, 19(3):373–403, 2011. doi:10.1162/EVCO_a_00030.

[168] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. *Proceedings of USENIX Security 2023*, 2023. doi:10.48550/arXiv.2302.04222.

[169] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. *Proceedings of IEEE S&P 2024*, 2024. doi:10.48550/arXiv.2310.13828.

[170] J. Shelley. The concept of the aesthetic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2022 edition, 2022. URL `https://plato.stanford.edu/entries/aesthetic-concept/`.

[171] H. Shevlin and M. Halina. Apply rich psychological terms in ai with care. *Nature Machine Intelligence*, 1:165–167, 2019. doi:10.1038/s42256-019-0039-y.

[172] B. Shu, A. Singh, and M. ElSherief. From syntax to emotion: A mechanistic analysis of emotion inference in llms. *arXiv preprint*, 2026. doi:10.48550/arXiv.2604.25866.

[173] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024. doi:10.1038/s41586-024-07566-y.

[174] R. Shusterman. *Body Consciousness: A Philosophy of Mindfulness and Somaesthetics*. Cambridge University Press, 2008. ISBN 978-0-521-67578-3. doi:10.1017/CBO9780511802829.

[175] F. N. Sibley. Aesthetic concepts. *The Philosophical Review*, 68(4):421–450, 1959. doi:10.4324/9781315303673-101.

[176] N. Singh, M. Cherep, and P. Maes. Discovering and steering interpretable concepts in large generative music models. *arXiv preprint*, 2025. doi:10.48550/arXiv.2505.18186.

[177] J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. Defining and characterizing reward hacking. *Proceedings of NeurIPS 2022*, 2022. doi:10.48550/arXiv.2209.13085.

[178] M. Skov and M. Nadal. There are no aesthetic emotions: Comment on Menninghaus et al. (2019). *Psychological Review*, 127(4):640–649, 2020. doi:10.1037/rev0000187.

[179] N. Sofroniew, I. Kauvar, W. Saunders, R. Chen, T. Henighan, S. Hydrie, C. Citro, A. Pearce, J. Tarng, W. Gurnee, J. Batson, S. Zimmerman, K. Rivoire, K. Fish, C. Olah, and J. Lindsey. Emotion concepts and their function in a large language model. *Transformer Circuits*, 2026. doi:10.48550/arXiv.2604.07729. arXiv:2604.07729; Anthropic Research, April 2026.

[180] L. B. Soros, A. Adams, S. Kalonaris, O. Witkowski, and C. Guckelsberger. Open-endedness is essential for artificial superhuman intelligence. *Proceedings of ICML 2024*, 2024. doi:10.48550/arXiv.2406.04268.

[181] K. O. Stanley and J. Lehman. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer, 2015. doi:10.1007/978-3-319-15524-1.

[182] L. Stark and J. Hoey. The ethics of emotion in artificial intelligence systems. In *Proceedings of FAccT 2021*, pages 782–793, 2021. doi:10.1145/3442188.3445939.

[183] P. Sui, Y. Zhu, T. Cheng, P. West, R. J. So, H. Long, and A. Holtzman. Spoiler alert: Narrative forecasting as a metric for tension in LLM storytelling. *arXiv preprint*, 2026. doi:10.48550/arXiv.2604.09854.

[184] L. Sun, L. Yan, X. Lu, A. Lee, J. Zhang, and J. Shao. Valence-arousal subspace in llms: Circular emotion geometry and multi-behavioral control. *arXiv preprint*, 2026. doi:10.48550/arXiv.2604.03147.

[185] V. Surkov, C. Wendler, A. Mari, M. Terekhov, J. Deschenaux, R. West, C. Gulcehre, and D. Bau. Unpacking sdxl turbo: Interpreting text-to-image models with sparse autoencoders. *Proceedings of NeurIPS 2025*, 2025. doi:10.48550/arXiv.2410.22366.

[186] P. Sweetser and P. Wyeth. Gameflow: A model for evaluating player enjoyment in games. *Computers in Entertainment*, 3(3):1–24, 2005. doi:10.1145/1077246.1077253.

[187] J. Taylor, W. Agnew, M. Sap, S. E. Fox, and H. Zhu. The algorithmic gaze of image quality assessment: An audit and trace ethnography of the laion-aesthetics predictor. *Proceedings of FAccT 2026*, 2026. doi:10.48550/arXiv.2601.09896.

[188] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, A. Tamkin, E. Durmus, T. Hume, F. Mosconi, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and D. Amodei. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits*, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/`. Anthropic Research, May 2024.

[189] L. Tětková, T. Brüsch, T. Dorszewski, F. M. Mager, R. Ø. Aagaard, J. Foldager, T. S. Alstrøm, and L. K. Hansen. On convex decision regions in deep network representations. *Nature Communications*, 16:803, 2025. doi:10.1038/s41467-025-60809-y.

[190] C. Tigges, O. J. Hollinsworth, A. Geiger, and N. Nanda. Linear representations of sentiment in large language models. *Proceedings of BlackboxNLP @ EMNLP 2024*, 2024. doi:10.48550/arXiv.2310.15154.

[191] G. Tuckute, A. Sathe, S. Srikant, M. Taliaferro, M. Wang, M. Schrimpf, K. Kay, and E. Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024. doi:10.1038/s41562-023-01783-7.

[192] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering language models with activation engineering. *arXiv preprint*, 2023. doi:10.48550/arXiv.2308.10248.

[193] S. Van de Cruys and J. Wagemans. Putting reward in art: A tentative prediction-error account of visual art. *i-Perception*, 2:1035–1062, 2011. doi:10.1068/i0466aap.

[194] O. Vartanian and A. Chatterjee. *Brain, Beauty, and Art*. Oxford University Press, 2022. ISBN 978-0-19-751362-0. doi:10.1093/oso/9780197513620.001.0001.

[195] E. A. Vessel, A. I. Isik, A. M. Belfi, J. L. Stahl, and G. G. Starr. The default-mode network represents aesthetic appeal that generalizes across visual domains. *Proceedings of the National Academy of Sciences*, 116(38):19155–19164, 2019. doi:10.1073/pnas.1902650116.

[196] K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. *Proceedings of ICLR 2023*, 2023. doi:10.48550/arXiv.2211.00593.

[197] S. Wang, S. Tan, S. Liu, H. Jia, G. Huang, J. Bailey, and T. Dang. Composable and controllable human-like emotional tts via activation steering. *arXiv preprint*, 2026. doi:10.48550/arXiv.2602.03420.

[198] E. Wassiliwizky, T. Jacobsen, J. Heinrich, M. Schneiderbauer, and W. Menninghaus. Tears falling on goosebumps: Co-occurrence of emotional lacrimation and emotional piloerection indicates a psychophysiological climax in emotional arousal. *Frontiers in Psychology*, 8:41, 2017. doi:10.3389/fpsyg.2017.00041.

[199] G. A. Wiggins. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems*, 19:449–458, 2006. doi:10.1016/j.knosys.2006.04.009.

[200] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image generation. *Proceedings of ICCV 2023*, 2023. doi:10.48550/arXiv.2306.09341.

[201] D. Wurgaft, C. Rager, M. Kowal, V. Shyam, S. Feucht, U. Bhalla, T. Haklay, E. Bigelow, R. Sarfati, T. McGrath, O. Lewis, J. Merullo, N. D. Goodman, T. Fel, A. Geiger, and E. S. Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. *arXiv preprint*, 2026. doi:10.48550/arXiv.2605.05115.

[202] D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. doi:10.1038/nn.4244.

[203] G. N. Yannakakis and J. Hallam. Towards optimizing entertainment in computer games. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(2):160–169, 2007. URL `https://yannakakis.net/wp-content/uploads/2012/02/AAI.pdf`.

[204] G. N. Yannakakis and J. Togelius. Experience-driven procedural content generation. *IEEE Transactions on Affective Computing*, 2(3):147–161, 2011. doi:10.1109/T-AFFC.2011.6.

[205] G. N. Yannakakis and J. Togelius. *Artificial Intelligence and Games*. Springer, 2018. ISBN 978-3-319-63519-4. doi:10.1007/978-3-319-63519-4.

[206] V. Zaigrajew, H. Baniecki, and P. Biecek. Interpreting clip with hierarchical sparse autoencoders. *Proceedings of ICML 2025*, 2025. doi:10.48550/arXiv.2502.20578.

[207] N. Zangwill. Aesthetic judgment. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2023 edition, 2023. URL `https://plato.stanford.edu/entries/aesthetic-judgment/`.

[208] S. Zeki, J. P. Romaya, D. M. T. Benincasa, and M. F. Atiyah. The experience of mathematical beauty and its neural correlates. *Frontiers in Human Neuroscience*, 8:68, 2014. doi:10.3389/fnhum.2014.00068.

[209] J. Zhang, J. Lehman, K. O. Stanley, and J. Clune. OMNI: Open-endedness via models of human notions of interestingness. *Proceedings of ICLR 2024*, 2024. doi:10.48550/arXiv.2306.01711.

[210] S. Zhang, W. Shi, S. Li, J. Liao, H. Cai, and X. Wang. Interpretable reward model via sparse autoencoder. *arXiv preprint*, 2025. doi:10.48550/arXiv.2508.08746.

[211] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint*, 2023. doi:10.48550/arXiv.2310.01405.