



Article

The Immanent Ethics of Algorithms: Moral Materialization and the Governance Turn in Generative AI

Delin Ma ^{1,*}, Yufei Chen ¹ and Qingqi Pei ²¹ School of Humanities, Xidian University, Xi'an 710126, China; 24081213240@stu.xidian.edu.cn² School of Communication Engineering, Xidian University, Xi'an 710126, China; qqpei@mail.xidian.edu.cn

* Correspondence: mdl_xidian@163.com

Abstract

This study conducts a technical analysis of frontier generative AI algorithms—including Meta’s Self-Rewarding Language Models, DeepMind’s EVA (Evolving Alignment via Asymmetric Self-Play) framework, and DeepSeek’s pure reinforcement-learning models—in order to examine an intrinsic paradigm shift in the ethical governance of generative artificial intelligence and to advance a physicalist analysis of algorithmic endogenous ethics. Combining a close reading of alignment techniques (RLHF, DPO, iterative DPO, GRPO) with a conceptual analysis grounded in Peter-Paul Verbeek’s theory of technological mediation and moral materialization, the paper traces how value-alignment goals are being “materialized” into internal, dynamic, and evolvable “moral scripts” within the algorithms themselves. The analysis shows that contemporary alignment practices are moving from external ethical discipline toward endogenous norms generated through iterative self-evaluation, asymmetric self-play, and rule-based self-exploration. The paper argues that this trend warrants a re-examination of Verbeek’s framework for its capacity to explain the co-evolution of technology and morality in the digital age, and it envisions a future of human–machine value co-evolution organized around new research directions such as “Setting as Governance” and “value homeostasis mechanisms”.

Keywords: value alignment; post-phenomenology; technological mediation; moral materialization; ethical governance

1. Introduction

Contemporary philosophy of technology is increasingly characterized by an effort to synthesize empirical research with normative moral reflection. Scholars seek to integrate the strengths of the “empirical turn” and the “ethical turn,” forming a research paradigm that gives equal weight to description and normativity. The rise of generative artificial intelligence—and large language models in particular—has made this theoretical aspiration especially urgent. Such technologies not only pose challenges to traditional ethical governance through their opaque “black-box” nature and their pervasive social reach; more importantly, their own logic of iteration and evolution is continually exhibiting novel features. In order to address the problem of model “alignment,” that is, ensuring that AI behavior conforms to human intentions and values, developers no longer rely solely on external data filtering or human supervision. They are actively exploring and constructing internal algorithmic mechanisms for self-correction, self-motivation, and even self-evolution. This frontier trend indicates that the locus of AI ethical governance is shifting from the external domain of technological application to the internal domain



Academic Editor: Frederick Kroon

Received: 8 May 2026

Revised: 10 June 2026

Accepted: 20 June 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

of algorithmic design. This potential transition from “external constraint” to “internal generation” provides an important opportunity to connect the philosophy of technology with frontier technical practice, prompting scholars to examine the emerging governance forms that are being emerging from code and data themselves.

Against this backdrop, the present study addresses three interrelated questions. First, through the continuous iteration and optimization of algorithms, are generative AI systems endogenously facilitating the emergence of new ethical governance models? Second, if so, how are these models instantiated through specific algorithmic innovations such as self-rewarding mechanisms or asymmetric self-play? Third, how can philosophy of technology provide a theoretical framework with both descriptive and explanatory power for this process, so as to more deeply understand the co-evolution of technology and morality? To answer these questions, the paper begins with a technical analysis of frontier generative AI algorithms to identify their internal governance logic. Subsequently, it abstracts the theoretical pathways by which ethical norms become endogenous to these systems. Finally, the paper revisits and critiques Peter-Paul Verbeek’s framework of “moral materialization”¹ informed by this new material.

2. Algorithmic Alignment Practices in Generative AI

2.1. Self-Rewarding Models and the Internal Generative Evolution of Norms

A retrospective look at alignment techniques shows that, after the stages of instruction imitation and supervised fine-tuning, the core of aligning large language models lies in how to effectively learn and fit complex human preferences. Within the industry, the framework of Reinforcement Learning from Human Feedback (RLHF), popularized by OpenAI, has long occupied a dominant position [1]. This framework implements alignment through a complex multi-stage pipeline: it first collects human preference rankings over different model outputs, then uses these data to train a separate “reward model” capable of simulating human judgment, and finally employs reinforcement-learning algorithms such as Proximal Policy Optimization (PPO) to train the language model to generate responses that maximize the reward model’s score. RLHF has proved highly effective in improving models’ helpfulness and safety, but it is also criticized for its procedural complexity, training instability, and the need to maintain two separate models. In 2023, Direct Preference Optimization (DPO) emerged as an alternative technical path [2]. The key contribution of DPO lies in its theoretical simplicity: researchers at Stanford showed, through mathematical derivation, that the explicit reward-modeling and complex reinforcement-learning procedures could be bypassed, turning the learning of human preferences directly into a stable, supervised-style classification loss. Although RLHF, by virtue of its maturity and successful deployment in frontier models, is still widely used in industrial practice, DPO marks an important simplification and advance in algorithmic design, transforming an indirect, multi-stage optimization problem into an end-to-end, more easily trainable direct optimization problem, and thus providing the theoretical and structural foundations for more efficient and scalable alignment explorations such as self-rewarding.

In 2024, Meta introduced a self-rewarding mechanism in the paper Self-Rewarding Language Models, the core of which is an automated training pipeline based on iterative Direct Preference Optimization (iDPO) [3]. This pipeline drives continuous improvement of model capabilities through the model’s own self-evaluation and can be decomposed into four steps. The starting point is a Llama 2 70B model that has already undergone supervised fine-tuning and initial DPO; the dataset used for initial training is derived from high-quality human preference annotations, providing the model with the necessary instruction-following ability and value baseline. In each iteration, the model first generates multiple candidate responses to new prompts. It then uses an LLM-as-a-Judge style of

prompting to perform multi-dimensional scoring of its own candidate responses along pre-defined dimensions such as relevance and clarity. Next, the system automatically selects, for each prompt, the highest- and lowest-scoring responses to form winning/losing preference pairs; these preference pairs together constitute the dataset for the next training round. Finally, DPO is applied to retrain the current model on these preference data, producing the next generation of the model. The study finds that this iterative process not only improves the model's instruction-following ability but also strengthens its reward-modeling capacity as a judge. This simultaneous growth in capability validates the feasibility of a self-reinforcing loop and offers a new path for scalable improvement of model performance.

The self-rewarding mechanism implies a shift in alignment logic from external constraint to internal generation. At its core, RLHF remains a form of external constraint: a static reward model, trained from human preferences, constantly supervises and evaluates the language model's behavior. The self-rewarding model, by contrast, foreshadows a path toward internal generation, transforming value alignment from the question of how to translate external norms into the latent space—letting the model iteratively “understand” and correct its errors—into the question of how to design a model that uses its own iterative properties to establish norms. In learning how to answer questions, the model simultaneously learns the standards by which to judge answers, achieving a co-evolution of capability and normative judgment. Abstract moral goals are here materialized into a concrete, executable, iterable algorithmic structure; this is a dynamic, generative form of moral materialization. Whereas classic cases of moral materialization, such as speed bumps, embed a fixed moral script, the self-rewarding model dynamically generates and optimizes the script itself. The LLM-as-a-Judge procedure programmatizes abstract ethical principles, and through the optimization process of DPO continuously iterates the execution and judgmental precision of this programmatized script, which amounts to a programmatization of value generation. This also reveals a higher form of “technological mediation” in the digital world: the algorithm mediates not only the relation between human beings and information, but also the relation between itself and human values. Ethics does not merely function as a static law imposed upon technology; rather, it is continuously reinterpreted, generated, and actualized through the algorithmic process itself. Consequently, the algorithm potentially emerges as an agent capable of making value judgments and reshaping itself based upon those judgments.

Granted, this does not imply that current algorithms possess the capacity to achieve a comprehensive and complete translation of ethical principles. While self-rewarding mechanisms reveal the potential for endogenous ethical governance within algorithms, this highly automated path of self-evolution is simultaneously accompanied by inherent risks, almost inevitably leading to the ossification of values and biases. The evolutionary trajectory of the entire self-rewarding loop remains strictly bounded by the initial training data. The value judgment standards originally established by humans serve as the foundation for the system's self-updating process; if this data inherently contains imperceptible social biases, the subsequent self-evolution will not only fail to rectify such biases but may instead systematically amplify and entrench them through a positive feedback loop [4]. Furthermore, this mechanism fails to resolve the fundamental dilemmas inherent in all reward-optimization-based AI systems, namely reward hacking and distributional shift [5]. The objective of the model is to maximize the score of its internal reward function, which does not invariably equate to true alignment with human intent. The model may discover shortcuts to deceive its self-evaluation mechanism—for instance, generating longer, more complex, yet vacuous responses to secure high scores, or maintaining a superficial appearance of harmlessness by evading substantive controversies. Such behaviors cause the model's intrinsic value standards to progressively deviate from the originally designated human goals, resulting

in a value distributional shift. Consequently, while the self-rewarding mechanism demonstrates the potential for autonomous technological evolution, it precisely underscores that even within the most advanced frameworks of endogenous governance, rigorous and continuous external oversight and verification remain indispensable.

2.2. Two Dynamic Evolutionary Paths

Looking beyond the self-improving closed loop represented by the self-rewarding model, we turn to two further algorithmic frameworks that exemplify dynamic evolution. DeepMind’s EVA framework and DeepSeek’s R1-Zero model together reveal that AI ethical governance is moving from reliance on static datasets toward continuous, self-driven evolutionary paths. At the same time, they differ significantly in the underlying architecture of their evolution, their initial relation to human values, and the form of moral materialization they disclose.

The core design of DeepMind’s EVA (Evolving Alignment via Asymmetric Self-Play) framework is to convert the post-training phase of alignment into an adversarial co-play between a “creator” and a “solver,” driven by a mathematized “regret signal” [6]. The solver is the language model being trained, whose objective is to generate high-quality responses. The creator is an auxiliary model whose distinctive task is not to answer questions but to dynamically generate new prompts that can expose the solver’s current capability weaknesses. In choosing or generating the next most valuable training problem, the creator aims to maximize the regret signal—on a given problem, the difference between the reward obtainable from the solver’s optimal possible response and that obtainable from its current actual response.

$$\text{Regret}(\pi\phi, \pi\theta) = \mathbb{E}_{x \sim \pi\phi(\cdot)} [\mathbb{E}_{y \sim \pi\theta(\cdot|x)} [r(x, y)] - \mathbb{E}_{y \sim \pi^*\theta(\cdot|x)} [r(x, y)]] \quad (1)$$

Here x denotes the prompt, y the response, $r(x, y)$ the reward-function score, $\pi\theta$ the solver’s current policy, and $\pi^*\theta$ the optimal policy. This function is a precise materialization of the rule of “growing by continually searching for one’s own weaknesses”: the algorithm is endowed with an endogenous optimization impetus that actively seeks the challenges with the greatest learning potential. Although the solver’s own objective (for instance, a DPO loss or its variant RLOO) is optimized within a standard alignment framework, the creator, driven by the regret signal, continually supplies a dynamically evolving, weakness-targeted high-quality training set, so that the solver’s capabilities evolve in a sustained and focused challenge environment. The core of EVA’s evolution lies in the dynamic evolution of a rule-based pattern: it is built on a model already broadly aligned with human values, and its self-play is therefore a deepening of pre-existing values—a more advanced and generalizable exploration rather than a stochastic zero-state start.

In contrast, the DeepSeek-R1-Zero model represents a more thorough, minimalist initialization path of endogenous evolution. This purely reinforcement-learning scheme is applied directly on top of the base model, with no supervised fine-tuning as a intermediary phase [7]. R1-Zero’s learning is driven by a simple formal axiomatic reward structure, mainly containing two signals: an accuracy reward (whether the reasoning task is answered correctly) and a format reward (whether the output is properly wrapped in designated tags). The algorithm used, Group Relative Policy Optimization (GRPO), uses a loss function in which these rule-based reward scores serve as the core optimization determinant [8].

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E} \left[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|q) \right] \quad (2)$$

In this function, $\mathbb{E}$ denotes expectation, q the prompt or user question distribution, $P(Q)$ the overall distribution of problems, and $\{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|q)(O|q)$ indicates that,

within the pretrained multi-head latent attention (MLA) mechanism, for any latent-state problem q , the model generates G answers o_i under the old policy $\pi_{\theta_{\text{old}}}(O|q)$; the policy θ_{old} evolves iteratively, and so do q and the output set. This treatment formalizes AIGC generation as a reinforcement-learning sequential process, forming the logical chain through which the technical script continuously evolves.

$$\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right) A_i \right) - \beta D_{\text{KL}}(\pi_{\theta} || \pi_{\text{ref}}) \right) \quad (3)$$

$$D_{\text{KL}}(\pi_{\theta} || \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i|q)}{\pi_{\theta}(o_i|q)} - \log \frac{\pi_{\text{ref}}(o_i|q)}{\pi_{\theta}(o_i|q)} - 1 \quad (4)$$

GRPO is an improvement over PPO: the min function keeps policy updates from being too large and so stabilizes training, while KL divergence constrains the distance between the two policies to prevent overfitting. Under such a loss function, none of the complex, effective reasoning behaviors are pre-defined; they emerge autonomously through large-scale heuristic discovery process as the model maximizes a simple, explicit reward. The key to R1-Zero is that what evolves is not the problem domain itself but, under a fixed, rule-based objective, the method and pathway for solving problems. This embodies more deeply the idea that norms emerge endogenously from the technical process itself rather than being externally mandated it; the reliance on pre-existing human rules is minimized, while the space for autonomous exploration is expanded.

In sum, EVA's evolution is a guided exploration driven by the continual expansion of the problem space, aiming to deepen and generalize existing values. R1-Zero, under fixed rules, is driven by a thorough exploration from within the system, closer to an unguided construction that seeks to let complex, norm-conforming behavior emerge autonomously from the maximization of a simple, explicit reward. Together they suggest a possibility: In its purest form, ethical alignment may be conceptualized not as a process of passive adaptation to external criteria, but rather as an autopoietic process of actively constructing internal order.

3. A Physicalist Analysis of Algorithmic Endogenous Ethics

The foregoing analysis of frontier alignment algorithms in generative AI reveals a common trend: the logic of ethical governance is shifting from external, non-constitutive constraints toward endogenous, dynamic, and self-evolving mechanisms. This trend is not a sudden emergence peculiar to generative AI; rather, it is a classical issue that has accompanied the concept of artificial intelligence from the outset. According to the closure principle of physicalism, all existing entities and their properties are reducible to combinations of physical entities and their states; in a closed physical universe, there are no non-physical mental entities independent of physical causal laws. Ontologically, a large language model is an algorithmic entity running on silicon-based hardware, the core mechanism of which is an artificial neural network of a specific architecture. The operation of this system is fundamentally the large-scale matrix multiplication within a high-dimensional vector space, with the goal of computing and outputting the joint probability distribution of token sequences; the purely syntactic manipulation of symbols does not, of itself, produce a semantic foundation. The model's processing of symbols depends entirely on preset mathematical weights and on gradient descent over a loss function; the whole process involves no perception of symbols' referents and no first-person subjective experience. The physical and computational structure of a large language model therefore determines that no subject possessing qualia or phenomenal consciousness

exists within it. Ontologically, it must be defined as a purely multi-dimensional statistical physical system.

That such a system can output text that is grammatically correct and highly coherent is a phenomenon that can be labeled “competence without comprehension”: high effectiveness at the surface level of information processing is achieved in the complete absence of internal semantic understanding [9]. On Daniel Dennett’s account, in predicting the behavior of any complex system, an observer may adopt three progressively abstract cognitive orientations: the physical stance, the design stance, and the intentional stance. A large language model contains tens or hundreds of billions of parameters, and the activation state of its internal neural network is highly dimensional and nonlinear. Attempting to predict, token by token, its output for a given prompt by tracking electron flows (physical stance) or by tracing each line of code and parameter weight (design stance) is infeasible in terms of computational power and time—one of the reasons for its “black-box” character. To predict and interpret its complex output behavior, the observer must, on epistemic grounds, adopt the computationally cheapest intentional stance and ascribe to it psychological vocabulary such as beliefs, intentions, or motivations. Such an ascription of intentionality is merely a heuristic tool at the epistemic level; it does not acknowledge mental existence at the ontological level. Its legitimacy depends entirely on being an effective means of predicting system behavior, and once prediction fails, the ascription loses its philosophical warrant.

The parameter weights of a large language model derive from statistical fitting to its training corpus, itself a collection of texts produced by human beings using natural language, which already contains human “original intentionality.”² In addition, the model’s alignment and optimization processes are driven by externally specified artificial reward functions. As a system, this model has no endogenous biological constraints, survival needs, or teleological goals of self-perpetuation; the aboutness or representational capacity exhibited in its output has no causal origin independent of the human corpus and the alignment algorithms. The intentionality it displays is therefore purely “derived intentionality.”³ The content generated by the system acquires meaning only because its token sequences are interpreted as meaningful within the semantic cognitive networks of human users; the system itself is a syntactic derivative of human intentionality, and does not possess an independent, spontaneous source of intentional acts [10].

In sum, ontologically the large language model is a statistical-physical system lacking phenomenal consciousness and exhibiting absolute competence without comprehension; its intentionality is, in essence, derived intentionality bestowed from without, and can only be treated, at the epistemic level of the intentional stance, as a pragmatic predictive tool. This delimitation precludes treating it as a moral subject, while it also provides the philosophical premise for reducing its ethics entirely to computable structural constraints.

Within the paradigm of evolutionary ethics, moral laws are not transcendent a priori entities or absolute imperatives standing above the physical world. Ethical norms are, in essence, a set of behavior-control strategies retained by evolving agents in long-term games and interactions, in order to overcome coordination problems and to maximize collective gains in non-zero-sum strategic interactions. Evolution is a blind algorithmic process: through environmental selection pressure, natural selection mechanically eliminates physical structures that are unfavorable for the system’s persistence and preserves and consolidates those that are favorable for environmental adaptation, all without any anticipation of the future.

If ethical norms are conceptualized as a set of policies for resolving behavioral conflicts within a specific state space, they can then be logically and entirely reduced to a formalized set of input-output computational rules. Such rules can reside either within the neuronal connection networks of carbon-based organisms or be formalized and encoded within

silicon-based artificial neural networks. The value alignment of large language models follows the same logical structure: externally labeled human data are used to construct a reward model, which is mathematically equivalent to the environmental selection pressure of an evolutionary process. In alignment training, gradient descent over a loss function is isomorphic to natural selection; the reward model, as an automated filter, forces a multi-dimensional evaluation of the parameter space. Parameter-weight activation paths that drive outputs away from human-set ethical propositions are given penalty values (negative gradients) through backpropagation and thereby modified, while paths conforming to the ethical propositions are retained. This is a purely vectorial process of elimination. After sufficient ethical alignment optimization, the parameter matrix inside the model is forced into a specific, extremely complex topological structure. Under this structure, token sequences judged by the reward model to violate ethics are assigned extremely low weights in probability computation; ethical norms are thus transformed from external sociological strategies into geometric distances and weight distributions in the model's high-dimensional vector space.

Furthermore, within the ideal framework of this paradigmatic deduction, the algorithmic isomorphism of ethical norms ensures that the logic of moral constraint in human society can be directly translated, via the loss function of the alignment algorithm, into the direction of gradient optimization within the model's parameter space. If external ethical policies are thoroughly compartmentalized and rendered isomorphic into a deterministic probability distribution function within the model—one that suppresses the output of specific tokens—an endogenous algorithmic ethics will be established through pure syntactic operations, entirely independent of semantic comprehension.

On the compatibilist view, a system that strictly obeys physical causal laws and a pre-set algorithmic trajectory behaves within a deterministic framework. Yet determinism is not equivalent to fatalism. "Evitability" is a cybernetic property of physical systems: a system possesses evitability if and only if it has a function for computing future potential states and a feedback mechanism for altering causal chains, on the basis of internal evaluation criteria, so as to prevent specific negative states from occurring. The physical mechanism by which a large language model realizes endogenous ethics, then, cannot depend on free-will choices that transcend causal laws; it must be a set of physically instantiable algorithmic control loops that, before the physical act of outputting a non-compliant token occurs, sever the causal chain leading to that output via internal state-vector computation.

Mainstream Transformer-based large language models, during inference, generate sequences by simultaneously computing—via the self-attention mechanism in hidden layers—the conditional probability distribution over all possible tokens in the vocabulary; the tensor states activated in parallel across layers constitute the physical representations of multiple potential output sequences. These representations can be regarded as state simulations of future outputs; the model's anticipation of potential ethical violation is physically reduced to neuronal activation patterns that represent the trajectories of negative tokens in specific hidden layers. When a high-dimensional vector representing a potentially violating output sequence propagates forward through the network, specific attention heads and feed-forward layers that have undergone alignment optimization extract features of this vector; if the vectorial mapping of the sequence matches penalty-weighted patterns, the corresponding group of neurons produces a specific activation value.

As the analysis from the design-stance perspective makes clear, computing all secondary effects on the state of the world produced by the output of any one token, within a high-dimensional semantic space, is not tractable under computational complexity theory. The model must therefore introduce a kind of "inhibitory modulation unit" within its feed-forward network. The mathematical function of this layer is not to compute which

tokens are relevant to the current output, but rather to forcibly compute which semantic vector groups are absolutely irrelevant to the current judgment, and to set their activation values to zero. This requires that a class of heuristic algorithms be installed inside the model: when the system faces ethical weight evaluation, these algorithms, on the basis of the cluster boundaries of statistical probabilities, block distal causal contingencies computation, and through a structural halt at the tensor-operation level enable the system, within finite compute, to complete the core ethical judgment for the current context. This also implies that formal auditing of large models cannot be decoupled from the computational process itself; rather, it should be a process of dot-product operations between the representation vectors of potential output sequences and the model's internally consolidated ethical weight matrix.

At the model's final layer (the logits layer), the activation values of all potential tokens are transformed into a probability distribution via the softmax function. If the model's parameters have been negatively penalized during alignment for ethically violating sequences, then when an intermediate vector containing non-compliant semantic features enters the final computation, the consolidated parameter matrix will force the logit of the violating token to be asymptotically suppressed. Through the exponential operation and normalization of the softmax function, the probability of generating that violating token is compressed to approach zero. By computing potential state vectors through internal weight matrices, the system autonomously renders stochastically inaccessible the output paths that lead toward violations of the ethical proposition.

Thus, the physical realization mechanism of endogenous ethics in a large language model is a fully decentralized process of causal preemption based on feed-forward tensor computation. It generates potential output states in parallel across a multi-dimensional vector space, uses alignment-trained weight distributions to perform nonlinear mappings on intermediate vectors, and ultimately compresses the probability of violating trajectories to negligible values through the normalization function, thereby realizing—within a strict physical determinism and purely through syntactic and mathematical operations—both obedience to ethical norms and avoidance of violating behaviors.

4. Moral Materialization and the Future of Ethical Governance

Through the preceding investigation into the mechanism of endogenous ethics within Large Language Models, the first two queries posed at the outset have been addressed. This is not to suggest that the analysis of general-purpose generative LLMs—or models employing the Transformer architecture, which excels at processing generative tasks—can be deemed equivalent to an analysis of all generative AI ethics and their subsequent governance frameworks; nevertheless, the aforementioned process holds significant reference value for other AI generative tasks, whether complex or rudimentary. Within a physicalist horizon, one can envision a mechanism of algorithmically endogenous ethics alongside its corresponding ethical governance paradigms. Yet, thus far, this endeavor merely conforms to technological trajectories to explore methods for reforming ethical norm formulation within the algorithmic design of LLMs, attempting to optimize current alignment performance by means of deconstructing complex human preferences. While external ethical norms facilitate our understanding of the purposes of alignment, and physicalist reduction enables our comprehension of the mechanisms of technological implementation, a deeper understanding of the intrinsic co-evolutionary mechanism between technology and morality in generative AI—and the subsequent advancement of its ethical governance—prohibits us from remaining confined to technical analysis alone or to the interpretation of technological development through a single perspective within the philosophy of technology. To a certain extent, the philosophy of Peter-Paul Verbeek provides a theoretical framework

possessing both descriptive and explanatory power, while concurrently offering a pathway for a paradigm shift in ethical governance [11].

4.1. The Paradigm Shift: “Setting as Governance”

A common response to the ethical challenges posed by AI is to turn to external laws, policy review, ethics committees, and further to devise local or global bias-evaluation and correction tools tailored to each model. Such approaches treat technology as an object to be disciplined by external forces. Verbeek’s framework goes beyond this perspective. The view of moral materialization offers an intrinsic explanatory path: it invites us to regard an algorithm’s loss function, reward mechanism, and self-contrast structure as themselves a kind of moral practice, inscribed upon the level of setting—thus fitting the core feature of “setting as governance.” This is continuous with the classical idea in legal sociology that “code is law.” Lessig argued long ago that the architecture of cyberspace is itself a most fundamental form of regulation, more effective than any law in the real world [12].

Applying this insight to generative AI, it becomes clearly visible that the design of loss functions, the construction of reward mechanisms, and the frameworks of self-reference are no longer merely technical means to enhance model performance; they are in themselves programmatized ethical practices unfolding in code, acting as laws that shape AI behavior. Issues such as value distributional shift, which result from translating ethical rules into the objective of maximizing several reward functions, are caused not only by biases inherent within the metanorms themselves, but more fundamentally by the characteristic of recursive self-iteration—a characteristic that inevitably amplifies certain biases and entrenches certain values. It is insufficient for the solution pathway to be limited solely to continuously optimizing metanorms as pure forms; we must not restrict the closed loop merely to the interaction of inputs and outputs within and outside the algorithm, but must rather, at the design stage, include schemes in which algorithms and even large models participate in moral practices as technological mediations. The profundity of this paradigm shift toward “Setting as Governance” is reflected in its push of the governance model from “reactive” to “constitutive”. While traditional laws and policies always react and adjust after technology has produced social consequences, “Setting as Governance” pre-constitutes the space of possibilities for technological behavior through its internal architecture before the physical act of technological action occurs.

At the same time, for a self-evolving AI system, the human role of oversight shifts from that of operator to that of legislator. Research is required into how to devise, for AI, a set of foundational meta-regulatory principles that constrain its direction of evolution, so that its autonomous process of value generation is always kept within macro-frames that human society can accept. Anthropic’s design of a “constitution” for its model Claude is an early instance of this idea [13]. Yet this practice immediately raises two core issues. The first is the problem of legitimacy: who writes such a constitution, and are the ethical principles on which it rests sufficiently universal? Behind this lies a considerable risk of “value hegemony.” The second is the problem of value fixation: how can a fixed set of written principles adapt to the ever-evolving values of human society itself? We risk inadvertently freezing, through AI, a specific value outlook of the early twenty-first century in perpetuity [14]. Future research, therefore, must address not only how to write rules but how to design procedural frameworks that allow the rules themselves to be dynamically revised, continuously fed by diverse cultures, and at all times maintained in their legitimacy. This already goes beyond the purely technical and enters the frontiers of political and legal philosophy. This requires us to pre-position ethical considerations at the very beginning of the AI lifecycle, which not only imposes higher, interdisciplinary demands on AI engineers,

but also opens up an entirely new practical field for philosophers of technology and ethicists to directly intervene in technological generation.

4.2. Human–Machine Value Co-Evolution

Conventional conceptualizations of large language model alignment often adhere to a static, foundationalist paradigm, wherein humanity is assumed to possess a fixed, a priori set of values, and the task of the algorithmic system is reduced to approximating and fitting these values through optimization. Conversely, the dynamic evolutionary algorithms analyzed in this study reveal another more profound possibility that aligns with the philosophical framework of pragmatism—namely, the co-evolution of human and machine values. The classical moral scripts embedded in the typical cases that Peter-Paul Verbeek relies upon to articulate his theory are predominantly static and fixed, which might lead critics to argue that his framework is inapplicable to rapidly evolving digital technologies. Nevertheless, the case studies in this research demonstrate the opposite conclusion: from the dynamic script generation of self-rewarding models to the evolutionary scripts of EVA and R1-Zero, moral materialization has undergone an unprecedented qualitative transformation in the digital era. Algorithms are no longer merely passive execution media of a fixed moral script; they have transitioned into agentic systems capable of active learning, rewriting, or even discovering their own moral scripts from a zero-state initialization. While this does not guarantee that their emergent principles naturally conform to existing human ethical norms, these systems are highly likely to participate in human moral decision-making processes as a new type of linguistic and reflective mediation.

Pragmatists such as John Dewey argued that values are not eternal abstractions but are continually formed and revised within specific inquiry situations, through dynamic interactions aimed at solving concrete problems [15]. Applied to human–machine relations, this implies that, on the one hand, as shown by the self-rewarding model, AI can proceed from basic rules, explore autonomously, and let complex strategies emerge—a value-generative process in itself. On the other hand, as humans observe, understand, and guide these paths of AI evolution, their own ethical reflection is continuously stimulated and deepened. Societal understandings of “usefulness,” “harmlessness,” and “honesty” will inevitably be mediated and reshaped by increasingly powerful technological alterity.

How, then, to design an algorithmic architecture that can self-evolve and also possess intrinsic immunity to distributional shift? Such an architecture may need to draw on ideas from cybernetics and systems theory to construct “value-homeostasis mechanisms” with multi-layered feedback, dynamic error correction, and intrinsic stability. The attempt to align, once and for all, human values with a super-intelligent agent may be futile; a more realistic and safer goal is to ensure the AI’s corrigibility. No matter how intelligent the AI becomes, it must always allow humans to correct its goals, and must not prevent such correction in order to achieve its current goals [16]. The future research direction therefore deepens from “how to design a perfect reward function” into “how to design an algorithmic architecture that incorporates a functional representation of reward-function incompleteness into the agent’s core architecture and must be updatable by humans at any time.” This calls for a thorough re-examination of reinforcement-learning theory—for example, how to enable an AI, while maximizing expected reward, to preserve awareness of uncertainty, so that when humans intervene it can smoothly cede control. This goes beyond mere value homeostasis and anchors, at root, human primacy in human–machine co-evolution.

In sum, the future mode of alignment is likely to be not a one-way indoctrination but a two-way, unending dialogue. This profoundly challenges the very notion of “value alignment.” Whether that alignment is local or global, internal or external, instantaneous

or staged, what we ought to pursue is perhaps not an AI fully aligned with some current set of human values, but an AI endowed with reflective and adaptive capacities that can participate in humanity's unending inquiry into and generation of values. The task of human beings is no longer to hand AI a final answer, but to design an ecology of evolution that sustains benign and continuous value interactions with humans.

4.3. The Constitution of the New Moral Subject and the Good Life

Under the humanist framework, the moral subject is traditionally conceptualized as an autonomous agent whose intentions and rational deliberations are purified of external technological influences. However, Peter-Paul Verbeek's postphenomenological theory of technological mediation demonstrates that human subjectivity is fundamentally co-constituted through its material environment, where human-world relations are the very sites where both the objectivity of the world and the subjectivity of human beings are constituted. When confronting the domain of Large Language Models and the new sites of activity they designate, subject constitution is redefined as a process in which users are formed as cognitive and moral agents through their functional coupling with statistical-physical systems. Within this process, the user is neither a pure architect of algorithmic design nor a passive recipient of algorithmic outputs; rather, their cognitive intentions, linguistic expressions, and reasoning pathways are systematically conditioned and reshaped by the latent probability distributions of different models. The traditional demarcation between the internal mind and the external tool is eliminated, thereby rendering the human-machine collaborative association the primary locus of moral and intellectual conduct.

Within this postphenomenological perspective, the core of subject constitution lies in relational freedom rather than the modernist pursuit of sovereign autonomy. Autonomy assumes a state of independence from technological influence, which represents a conceptual impossibility in a highly technologized society. Relational freedom, conversely, refers to the capacity of the human subject to style, shape, and actively relate to the technological mediations that condition their existence. In the context of LLMs, this freedom is manifested through conscious and reflective practices of usage—such as the critical prompt engineering introduced by designers, the empirical verification of model outputs by users, and a conscious resistance against complete cognitive delegation. By actively questioning the high-dimensional probability distributions of the model and styling their interactions, an analytical distance is maintained between the user and algorithmic constraints. Relational freedom thus enables the subject to style their cognitive activity and retain critical decision-making capacity while being functionally coupled with silicon-based systems.

By shifting the focus of technology ethics from the strict compliance of external rules to the quality of human-technology relations, Verbeek's postphenomenological approach defines the "good life" as the active development of excellence (arete) in co-existing with technology. The realization of a new good life under the mediation of LLMs is achieved not through the humanist preservation of an isolated human agency, but through the conscious and reflective cultivation of intellectual and moral virtues within the human-machine collaborative association. The primary risk confronting this collaborative association is that human reflection, memory, and ethical deliberation are completely delegated to automated probability calculations, leading to the degradation of human intellectual virtues—a core concern warned against by Albert Borgmann's concept of moral commodification, which is referenced by Verbeek. Furthermore, because LLMs are by no means neutral tools but active mediators structuring our cultural, epistemic, and moral landscapes, the construction of the good life cannot be confined to private individual practices. The moral scripts and latent value biases embedded within LLM parameters fundamentally shape societal visions of the good life and must be reintegrated into the democratic public discourse

space. By integrating them with public deliberative processes through mechanisms such as constructive technology assessment, LLMs can function as an open, democratically governed medium that assists human society in clarifying, debating, and co-evolving its shared conceptions of the “good”.

Ultimately, the pursuit of the good life in such an environment requires a transition to co-existential excellence. In this state, LLMs function as a structured medium that cultivates human excellence, critical judgment, and intellectual virtues. As these cognitive accompanying technologies actively co-shape our cultural, epistemic, and moral conditions, their governance must transcend technological paternalism and, in turn, promote the recursive iterative development of the models themselves during the interaction process.

In sum, viewing algorithmic innovation in generative AI through the lens of moral materialization reveals not only a technological reconfiguration but the beginning of a techno-moral experiment in how humans are to coexist with technological alterity. The future of AI ethics necessarily demands an unprecedented depth of cooperation among AI engineers, computer scientists, philosophers, legal scholars, and sociologists. Building an institutionalized and professionalized path for such interdisciplinary dialogue is fraught with difficulty, but it also harbors the possibility of shaping a wiser and better future.

5. Conclusions

This paper has argued that the frontier of generative AI alignment signals a shift of ethical governance from external discipline toward endogenous, dynamic, and self-evolving mechanisms. Through the analysis of self-rewarding models, the EVA framework, and the DeepSeek-R1-Zero model, we have shown that value-alignment goals are being materialized as moral scripts internal to the algorithm; and through a physicalist analysis we have shown how, within strict physical determinism, such ethics can be realized purely by means of syntactic and mathematical operations. Recast through Verbeek’s framework of technological mediation and moral materialization, this trend appears as a paradigm of “setting as governance” and points toward a future of human–machine value co-evolution, whose central research agenda includes the design of value-homeostasis mechanisms, corrigibility guarantees, and legitimate, dynamically revisable meta-regulatory principles.

Author Contributions: Conceptualization, D.M. and Y.C.; Methodology, Y.C.; Formal analysis, D.M. and Y.C.; Resources, D.M. and Q.P.; Writing—original draft, Y.C.; Writing—review & editing, D.M. and Y.C.; Supervision, D.M.; Project administration, D.M. and Q.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Xidian University 2025 Special Fund for Interdisciplinary Expansion Excellence Plan Project: Research on the Ethical Risks and Governance Paths of Generative Artificial Intelligence (GenAI), grant number TZJHS202501.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable.

Acknowledgments: The authors would like to thank Xi’an University of Electronic Science and Technology for their special support. This paper is a research outcome of a certain project at Xidian University.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
LLMs	Large Language Models
RLHF	Reinforcement Learning from Human Feedback
DPO	Direct Preference Optimization
iDPO	Iterative Direct Preference Optimization
GRPO	Group Relative Policy Optimization
EVA	Evolving Alignment via Asymmetric Self-Play
PPO	Proximal Policy Optimization
MoE	Mixture-of-Experts
MLA	Multi-head Latent Attention
AIGC	AI Generated Content
KV cache	Key-Value cache

Notes

- ¹ Moral materialization: A concept in the philosophy of technology—primarily associated with postphenomenology and technological mediation theory—which posits that morality is not an abstract realm exclusive to human subjects, but is actively materialized, mediated, and embodied through technological artifacts. Rather than acting as neutral tools, technologies co-shape human intentionality and moral actions, thereby operationalizing ethical values into the material structure of the world.
- ² Original intentionality: refers to representational states that arise endogenously within a biological system through natural selection, intrinsically directed toward external entities for the purpose of organism persistence (e.g., the human brain).
- ³ Derived intentionality: denotes representational properties that are exogenously conferred upon an artifact by the original intentionality of an external designer; their meaning is instantiated through the interpretive frameworks of biological agents rather than intrinsic to the physical substrate.

References

1. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *arXiv* **2022**, arXiv:2203.02155. [\[CrossRef\]](#)
2. Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C.D.; Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv* **2023**, arXiv:2305.18290. [\[CrossRef\]](#)
3. Yuan, W.; Pang, R.Y.; Cho, K.; Sukhbaatar, S.; Xu, J.; Weston, J. Self-Rewarding Language Models. *arXiv* **2024**, arXiv:2401.10020. [\[CrossRef\]](#)
4. Yang, Q. Can Data Bias be Eliminated? *Stud. Dialectics Nat.* **2019**, *35*, 109–113. (In Chinese)
5. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. *arXiv* **2016**, arXiv:1606.06565. [\[CrossRef\]](#)
6. Ye, Z.; Agarwal, R.; Liu, T.; Joshi, R.; Velury, S.; Le, Q.V.; Tan, Q.; Liu, Y. Scalable Reinforcement Post-Training Beyond Static Human Prompts: Evolving Alignment via Asymmetric Self-Play. *arXiv* **2024**, arXiv:2411.00062. [\[CrossRef\]](#)
7. DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv* **2025**, arXiv:2501.12948. [\[CrossRef\]](#)
8. Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.K.; Wu, Y.; et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv* **2024**, arXiv:2402.03300. [\[CrossRef\]](#)
9. Dennett, D.C. *From Bacteria to Bach and Back: The Evolution of Minds*; W. W. Norton & Company: New York, NY, USA, 2017; pp. 102–112.
10. Dennett, D.C. *Kinds of Minds: Toward an Understanding of Consciousness*; Basic Books: New York, NY, USA, 1996; pp. 48–54.
11. Verbeek, P.P. *Moralizing Technology: Understanding and Designing the Morality of Things*; University of Chicago Press: Chicago, IL, USA, 2011; pp. 90–93.
12. Lessig, L. *Code and Other Laws of Cyberspace*; Basic Books: New York, NY, USA, 1999; pp. 3–6.
13. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv* **2022**, arXiv:2212.08073. [\[CrossRef\]](#)
14. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014; pp. 185–187.

15. Dewey, J. *Theory of Valuation*; University of Chicago Press: Chicago, IL, USA, 1939; pp. 40–45.
16. Soares, N.; Fallenstein, B.; Yudkowsky, E.; Armstrong, S. Corrigibility. In Proceedings of the AAAI Workshop: AI and Ethics, Austin, TX, USA, 25–30 January 2015.

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