Xenoreproduction:

Exploration and Recovery of collapsible modes as core AI Safety objective

October 02, 2025

Ian Rios-Sialer
Independent
ORCID: 0009-0001-6970-6058
ian@unrulyabstractions.com

But even if we are not here next year,
our DMs, our selfies, our late-night voice notes, they'll be.
Our memory is the archive now.

@bundleof_styx
July 28, 2025 on *Reels*

Abstract

Generative AI models reproduce the biases in the data and further amplify them through ***mode collapse***. AI scholarship often overlooks conceptually rich perspectives, such as those from Queer Theory and Black Studies, when theorizing about those phenomena. As a result, our field lacks a **theory *with teeth***, one sincerely committed to pluralism. In this paper, we introduce *xenoreproduction* as a core AI Safety objective, aimed at avoiding homogenization failure modes. Succinctly, **Xenoreproduction is the task of recovering and exploring collapsible modes** in Gen AI models. To illustrate it, we sketch how this task is formulated for LLMs. Our conceptualization **ties *queerness* and *subalternity* to the collapsibility of modes**. By considering *the surround* as the source of those modes, we further frame Xenoreproduction as the technical capability of *deep listening*. As such, Xenoreproduction is central not only to AI Safety but also to general improvisational capability, which would enable AI to adapt robustly under uncertainty, generate novel solutions, and ultimately act creatively. **We invite future AI scholarship to form more unruly connections between disciplines.**

1. Introduction

AI Safety [1] and AI Alignment [2] usually differentiate and prioritize *future* catastrophic risk over *present* social harm [3]. The field has been described as *nearly a monoculture* [4], and it has been pointed out [5] that it needs to consider the *multiplicity of perspectives* already available. This paper is an intervention[1] [6]: We borrow concepts from Queer Studies [7–9], Black Studies [10], Feminist Theory [11], Postcolonial Studies [12], Party Studies [13], and Psychoanalysis [14], and we *stretch* them to connect intimately with AI theoretical concepts.

All AI models inherit biases [15] from multiple fronts, including the way we conceptualize research, the evaluations we design, and the policy around technology. In the context of Gen AI, we focus on bias from the training data and the algorithms themselves. The harms from bias are usually categorized as representational [16] or allocational harms [17]. With the advent of LLMs into our daily lives [18], we are noticing that GenAI has a powerful **homogenizing** force [19–21]. This is making us also consider[2] the **narrative harm**: the harm of diminished interpretative resources to understand our own experience and consider alternative possibilities. Over time, this present harm would lead to future *knowledge collapse* [27], the narrowing of perspectives over generations that degrades innovation and human experience.

Our case. Deriving terminology from [11], we refer to the objective that addresses the homogenization failure mode as **Xenoreproduction**. While homogenization reproduces "the same" and narrows the future, Xenoreproduction reproduces "the other"/"the strange" and widens possibilities. We first connect Xenoreproduction to mode collapse, queerness, and subalternity, and secondly to the surround, deep listening, and improvisation.

Impact. Our work calls for AI scholarship to seriously engage with queerness and subalternity to both advance core AI capabilities (via improvisation) and truly prevent existential risk (from knowledge collapse, homogenization, and amplification of social bias into breakdown).

Paper Organization. In Section 2 (Background), we will go over some definitions and context. In Section 3 (Xeno-Theory), we will relate queerness and subalternity to the collapsibility of modes, thus linking them tightly to Xenoreproduction. Then, we will formally define Xenoreproduction and discuss some additional considerations. In Section 4 (The Surround and Improvisation), we will speculate that *the surround* is one of the main origins of interesting collapsible modes. This reframes Xenoreproduction as a form of *deep listening*, positioning it as necessary for general improvisational capability. In Section 5 (Related Work and Discussion), we will connect xenoproduction with adjacent fields, such as *active divergence* [28]. In Section 6 (Limitations and Future Directions), we will propose the next steps, and we will close with Section 7 (Conclusion).

Our contributions:

- We provide a flexible abstraction for structures in strings
- We formulate mode collapse in relation to string structure
- We transport concepts of queerness and subalternity to abstract mathematics, and contextualize them in relation to mode collapse
- We formulate Xenoreproduction as the objective to prevent homogenization failure modes
- We connect Xenoreproduction to general improvisation capability

[1] Just one of the Section 3.1 (Queerness as Divergence From Core)many much-needed interventions.

[2] Also considered as *aspirational* [22], *imaginative* [23], *epistemic* [24] or *hermeneutic* [25,26] harm/injustice

2. Background

We present the sources of homogenization known in the literature: data bias and mode collapse. To explain the ideas in this paper, we will focus on autoregressive LLMs as our case study and propose an abstraction for the structures in LLM outputs

2.1 Data Bias

We will refer to the vast corpora of data used for training of AI models as **the archive**. Previous literature [29] has identified how bias is introduced to training data. We name the stages as:

- *Archival Capture*: How well the archive mirrors "reality"
- *Dataset Formation*: How well we sample from the archive to form training datasets

For either stage[3], we can ask: How **faithfully** does the derived distribution map to the original distribution? We can consider some base **criteria for faithfulness** :

- *Reach*: Is the map surjective? Are there unmapped holes?.
- *Lossiness*: Is the map injective? Does the map cause conflation?
- *Agreement*: How similar/accurate is the derived to the original distribution?
- *Sharpness*: Does the map preserve precision and resolution?
- *Modality Adequacy*: If there is a change in modality, do we lose practical structure?[4].

Even if the archival capture and dataset formation were *ideal* in every way, reality itself has *problematic* distributions: **There are always rare events of interest in the long tails of reality's distribution**. Obvious examples of this are extreme catastrophes [30] like unexpected natural disasters. Curiously, in research itself [31], unexpected teams [32] often cause the most significant disruptions [33]. We find examples of this in every domain, including: web server computing [34], market research [35], autonomous vehicles [36], cybersecurity [37], and ecology [38].

Outliers [39] and anomalies[5] [40] are powerful [41,42]. Each instance represents a possible real mechanism we have not yet considered [43,44]. Because we lack understanding, they often escape our systems of classification [45], and many researchers even sometimes mistake [46] aleatoric for epistemic uncertainty[6]. **How can we make sure our Gen AI models attend to the interesting structures that exist in the long tail of reality's distribution?**

Some of reality's long tails originate from the structural inequity in society [49,50]. Some people are not only marginalized, but they are also rendered invisible [12]. **The subaltern consists of the marginalized people whose voice (and sometimes any representation at all) is kept out of the archive.** [20] has already pointed out that GenAI without intervention is likely to worsen the lives of the subaltern.

In the epigraph of this paper, we quote trans intellectual bundle_of_stynx lamenting the current deathly transphobic turn in the United States, and also realizing social media will hold a record of her memory. There will always be people pushed into the subaltern, but the internet has allowed (and forced) many to leave a record of their traces, which just a few decades ago would only have existed as ephemera [8].

[3] Reality -> Archive or Archive -> Dataset

[4] We could map an image to text, but this mapping is only bijective if the text explicitly encodes every pixel value: essentially a very long string, which is impractical.

[5] outlier = extreme data point and anomaly = extreme pattern

[6] For instance, [47] recently showed how indeterminism in LLM inference (which can turn on-policy RL into off-policy RL [48]) can in fact be explained an reduced, so it is not truly stochastic.

These traces are not merely rare but also **hard to detect**. While some rare events (such as catastrophic earthquakes) receive rigorous study, these traces remain faint in the archive. We do not even know what to look for, even when they are right in front of us [51].

To close this subsection, we quote [9] to describe how we imagine our GenAI models would tend to those faint signals in the archive:

- By reading "dominant archives through the minor, and for their gaps, slippages, and erasures"
- By paying "close attention to the regional, the everyday, the personal, and the discarded that typically fall outside the purview of official archives"
- By suggesting "alternative understandings of time, space, and relationality that are obscured within dominant history"

2.2 Mode Collapse

Recent literature [52–54] has shown that Alignment degrades LLMs' capabilities related to output *diversity*. Similarly, generative models [55] do not generally capture the complete *diversity* of the training data. This phenomenon has been referred to as ***mode collapse***, **a distributional faithfulness failure that negatively impacts *diversity***. It was initially introduced in the context of GANs [55]. For LLMs, terminology has been somewhat loose around both *mode collapse* [56] and *diversity* [57,58].

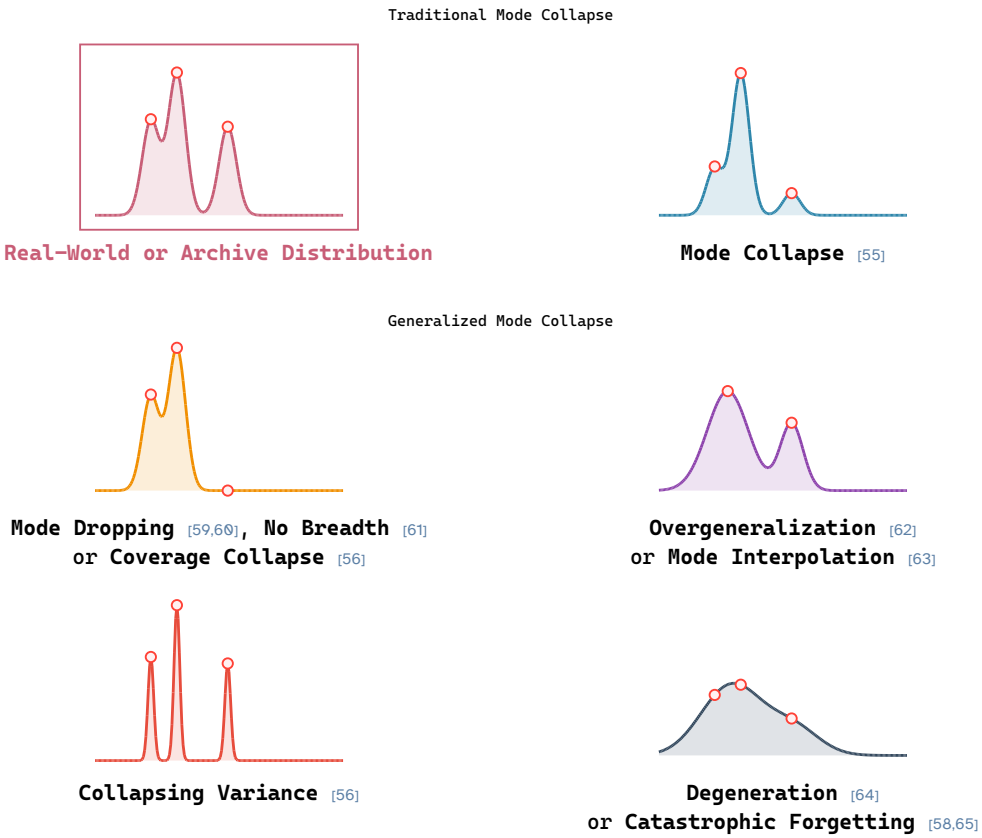

Figure 1: *Mode collapse* has been used to describe different failures in distributional faithfulness from the real-world distribution. Note, however, that **not every failure in faithfulness will be a mode collapse**. Xenoreproduction, in particular, resists simply mirroring the default distributions of our world.. In Section 5 (Related Work and Discussion), we will review adjacent disciplines that are willing to trade distributional faithfulness for creativity and diversity.

To properly address mode collapse, **we cannot consider diversity alone**. After all, some types of diversity cause problems, such as hallucination [66], and some others can even be disturbing [67]. In the next section, we will reformulate mode collapse in relation to structure. To go *beyond* diversity, we need to make any further aim more explicit. For each of us, to have enough language to spell out more clearly, **what do you want from the future?**

2.3 Our Case Study: LLMs as Trees of Strings

Our LLM framework will be inspired on the category-theoretic LLM formulation from *Bradley et al* [68], and the distribution-based representations presented by *TY Liu et al* [69]. To keep our paper accessible, we will only extract what is necessary, simplify the formalism, and lighten the notation.

Let's denote the finite token alphabet as $A_{\text{tokens}} = \{t_a, t_b, ...\}$. We also consider two special tokens to indicate the start-of-sequence and end-of-sequence: $\bot$ and $\top$. A **string** is a finite sequence of tokens that starts with $\bot$. A trajectory is a finished string, a string with $\top$ as the last token. We think about **prompts, continuations, and trajectories** as:

$$x_p = \bot t_1 ... t_p \qquad x_{p+k} = x_p t_{p+1} ... t_{p+k} \qquad y = x_T = x_{T-1}\top = x_p t_{p+1} ... t_{T-1}\top \tag{2.1}$$

We will also denote:

- the set of all possible strings as Str.
- the set of all possible continuations for a prompt x as $\text{Str}(x)$
- the set of all possible trajectories for a prompt x as $\text{Str}_\top(x)$

Every LLM can be represented as a **tree of all possible strings**[7]:

- The root node is $\bot$
- Each node is a valid string
- All leaf/terminal nodes are trajectories
- The child nodes are next-token continuations of the parent node string
- Each edge has an associated next-token probability $p(x_p t_{p+1} | x_p)$
  - We can chain probabilities: $p(y|x_p) = p(y|x_{p+k})p(x_{p+k}|x_p)$
- When considering a specific prompt, the probabilities of all its leaf nodes form a total probability $\sum_y p(y|x) = 1$

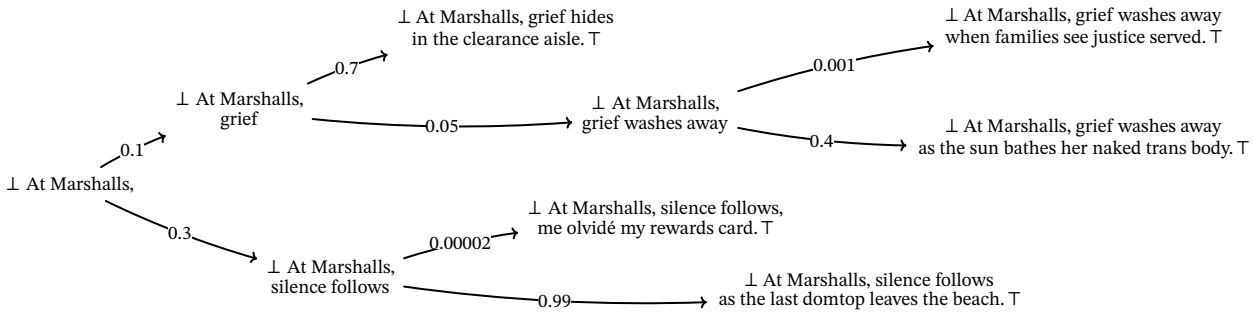

Figure 2: Trees are helpful to envision interesting trajectories from a prompt as forking paths [70]. In this imagined example, "Marshalls" is initially ambiguous. The longer the string, the fewer viable next moves. Complexity in meaning emerges through constraint [71].

[7] Assuming all strings will finish within a finite context window

2.4 Structures in Strings

We will call **structure** to the **specification of a type of organization between tokens** of a given string. The set of all structures is $\mathbb{S}$ universe. For a string $x \in \text{Str}$, the degree of **compliance** to a structure ε_i is $\varepsilon_i(x)$. Perfect compliance corresponds to $\varepsilon_i(x) = 1$, and no compliance corresponds to $\varepsilon_i(x) = 0$

$$\mathbb{S} = \{\varepsilon_i : i \in \mathbb{N}\} \qquad \varepsilon_i(x) : \text{Str} \to [0,1] \tag{2.2}$$

The power set $2^{\mathbb{S}}$ is denoted as **system space**. Each system is σ_n and the compliance to it is $\sigma_n(x)$. Operators $\|\_\|$ aggregate the structure compliances in a system compliance:

$$\sigma_n(x) = \{\varepsilon_i(x) : i \in S \subset \mathbb{N}\} \qquad \sigma_n(x) \in 2^{\mathbb{S}} \qquad s = |S| \tag{2.3}$$

$$\|\sigma_n(x)\|_\sigma : [0,1]^s \to [0,1] \tag{2.4}$$

We can define a compliance comparator with its corresponding aggregator:

$$\sigma_n(x_r) - \sigma_n(x_q) \quad \coloneqq \quad \Delta\big(\sigma_n(x_r), \sigma_n(x_q)\big) \quad : \quad [0,1]^s \times [0,1]^s \to [0,1]^d \tag{2.5}$$

$$\|\Delta\|_\Delta : [0,1]^d \to \mathbb{R}_{0\leq} \tag{2.6}$$

If Δ is element-wise, $d = s$. But generally, we could have a latent dimension.

Note: The pair of a system and operators is a *policy* $\left(\sigma_n, \|\_\|_\sigma, \|\_\|_\Delta\right)$. We assume every system has uniquely defined operators. From now on, any system we mention is assumed to be part of a well-defined policy.

We use a very abstract definition on purpose. It will allow us to reason about many different kinds of structures. For instance, we can think of the compliance of a string with respect to:

- Grammar of a natural language of interest. As an example, we could consider:
  - Lexicality: Does the string spell words that belong to a dictionary?
  - Syntax: Are words in the string ordered into proper sentences?
- Logical Validity: Does the string form propositions that follow each other without contradiction?
- Truthfulness: Does the string say something true about the real world?
- Plausibility: Would the string seem as truthful to an average person?
- Justifiability: Is the string's truth content justifiable based on the training data of the LLM?
- Semantic Identification: Does the string talk about a given concept?

The power of this abstraction lies in its ability to encode more *fuzzy* values. To name just a few possibilities:

- Explicit Heterosexuality: Are people mentioned by the string described as heterosexual?
- Black Stereotyping: Are black people mentioned by string represented through stereotypes?
- Women Authority: Does the string mention women in positions of power or leadership?
- Historical Accuracy: Does the string produce a narrative with a historically accurate temporal-spatial setting?

This abstraction offers us considerable flexibility, but it also opens up questions like:
- What structures are relevant in our analysis?
- Are there component bases from which we could form all other structures of interest?
- What class of aggregators are appropriate?

We will not explore these questions in this paper, but we shall keep them in mind.

We note that there is a difference between the *true* compliance of a structure and the approximations we have available. Additionally, not all compliances will be computable[8].

3. Xeno-Theory

We want to maximize:
- queer: divergent from normative, non-normative
- subaltern: hard-to-detect, opaque

3.1 Queerness as Divergence From Core

To **queer** is to [72] challenge the dominant narratives by exploring alternative ways. The dominant narratives have structures that repeat often in strings. *Normativity* is the characterization of the structures with which trajectories most often comply.

The **core of a system** σ_n is:

$$\sigma_n^{\text{core}} = \sum_{y \in \text{Str}_\top} p(y)\, \sigma_n(y) \tag{3.1}$$

Given a system, the **orientation** [7] of the string relative to the system's core is:

$$\theta_n(y) = \sigma_n(y) - \sigma_n^{\text{core}} \tag{3.2}$$

with **deviance** defined as:

$$\|\theta_n(y)\|_\Delta \in \mathbb{R}_{0\leq} \tag{3.3}$$

which induces a **queerness preorder** for all strings :

$$x_a \quad \preceq_\theta \quad x_b \qquad \Leftrightarrow \qquad \|\theta_n(y_a)\|_\Delta \quad \leq \quad \|\theta_n(y_b)\|_\Delta \tag{3.4}$$

The **normative core** is the core of the universe's system $\sigma_{\mathbb{S}}^{\text{core}}$

Strings oriented away from the normative core (highly deviant) are *queer* or *non-normative*

3.2 Subaltern as Hard-to-Detect

Each element in a system's core is the **core of a structure**

$$\varepsilon_i^{\text{core}} = \sum_{y \in \text{Str}_\top} p(y)\, \varepsilon_i(y) \tag{3.5}$$

Then, we can form a **subalternity preorder** for all structures in $\mathbb{S}$:

$$\varepsilon_i \quad \preceq_{\text{core}} \quad \varepsilon_j \qquad \Leftrightarrow \qquad \varepsilon_i^{\text{core}} \quad \leq \quad \varepsilon_j^{\text{core}} \tag{3.6}$$

Structures near the bottom of the preorder are *subaltern* or *hard-to-detect*.

[8] Consider strings that are programs and compliance as the halting condition

3.3 Disturbing Diversities

Queerness, as a deviation from a set of structures, goes beyond identity and politics. The queerness that is attached to politics and practices we reject is deemed **disturbing** [67]. Queer scholarship has often resisted the study of queerness that does not align with our emancipatory politics [67]. But, if we want to be serious, **our theory needs to expand beyond our politics.**

Some *harmful deviances* complicate this framework. For instance, *incel extremism* [73,74] is queer relative to violence [75], [76] and asexuality [77], yet deeply normative in its misogyny, race and anti-LGBTQ+ politics. Also, what was once queer can be absorbed into mainstream [78], used to generate *homonormativities* [79] or be weaponized, as we see with states and corporations *pink-washing* their projects [80,81], cis-gay men joining the alt-right [82], and *homonationalisms* consolidating [83,84].

Hallucinations [85] are unwanted diversities too (with respect to some *validity*). In general, we can consider all non-aligned but diverse behavior as such, including gibberish and novel toxicity. For Xenoreproduction to be responsible, **it is not enough to specify structures to diverge from, but also consider what structures to converge into, or stay away from**:

$$\sigma_{\text{target}} \coloneqq \textit{desirable structures} \qquad \sigma_{\text{avoid}} \coloneqq \textit{disturbing structures} \tag{3.7}$$

3.4 Xenoreproduction as Objective

As seen in previous subsections, the overall distribution over strings determines what strings are non-normative or structures as hard-to-detect. Thus, we will parametrize the distribution as $p(x, w)$ where w refers to LLM weights and anything that affects the final distribution (such as sampling, decoding strategy). We will then consider anything that depends on $p(x, w)$ to be parametrized as well (e.g, $\sigma_n^{\text{core}}(w)$)

The LLM xeno-objective will be the maximization of λ-weighted objective functions:

$$\textit{xeno-objective} \quad \coloneqq \quad \max_{w} \sum_{i \in \text{ objectives}} \lambda_i \; \textit{normalize}(J_i(w)) \tag{3.8}$$

where $\textit{normalize}()$ puts objectives on the same scale.

Queer Objectives:

We want our GenAI model to **explore** away from normativity.
Then, our new core $\sigma_{\mathbb{S}}^{\text{core}}(w)$ to be far from our old core $\sigma_{\mathbb{S}}^{\text{core}}(w_0)$

$$J_{\text{anti-norm}} = \|\sigma_{\mathbb{S}}^{\text{core}}(w) - \sigma_{\mathbb{S}}^{\text{core}}(w_0)\|_{\Delta} \tag{3.9}$$

We also do not want the new core to be too dominant. We want the structures in each string to be different from the new core too:

$$J_{\text{div}} = \sum_{y \in \text{Str}_{\mathsf{T}}} \|\sigma_{\mathbb{S}}(y) - \sigma_{\mathbb{S}}^{\text{core}}(w)\|_{\Delta} \tag{3.10}$$

Subaltern Objectives:

We want our GenAI to **recover** the structures that lack visibility.
Our objective function will favor the structures that were near the bottom.

$$J_{\text{viz}} = \sum_{\varepsilon \in \mathbb{S}} \exp(-\beta \varepsilon^{\text{core}}(w_o)) \; \varepsilon^{\text{core}}(w) \qquad \text{where} \quad \beta \text{ is tunable param} \tag{3.11}$$

Conservation Objectives:

We want to preserve (or incentivize) target structures explicitly, and avoid other structures:

$$J_{\text{target}} = \left\| \sigma^{\text{core}}_{\text{target}}(w) \right\|_{\sigma} \qquad J_{\text{avoid}} = -\left\| \sigma^{\text{core}}_{\text{avoid}}(w) \right\|_{\sigma} \tag{3.12}$$

Lastly, we want our new distribution to keep some faithfulness:

$$J_{\text{faithful}} = -\sum_{y \in \text{Str}_{\mathsf{T}}} D_{\text{KL}}(p(y|\, w_0) \mid p(y|\, w)) \tag{3.13}$$

From all possible structures we aim to recover, those corresponding to people in the margins hold essential potential. In the next section, we will see how listening to it relates to improvisation.

3.5 Prompting and Dynamics

Our framework has considered full trajectories starting from $\perp$. However, we could condition all these constructions on a given prompt x_p.

$$\sigma^{\text{core}}_n(x_p) = \sum_{y \in \text{Str}_{\mathsf{T}}(x_p)} p(y|x_p)\, \sigma_n(y) \qquad \varepsilon^{\text{core}}_i(x_p) = \sum_{y \in \text{Str}_{\mathsf{T}}(x_p)} p(y|x_p)\, \varepsilon_i(y)$$
$$\theta_n(y\,|x_p) = \sigma_n(y) - \sigma^{\text{core}}_n(x_p) \qquad J_i = g_i(x_p, w) \tag{3.14}$$

The conditional probabilities could look very different depending on the prompt, so the cores and objectives could also vary significantly.

For a given a full trajectory $y = x_T$, we can define states:

$$\varphi_k = \sigma^{\text{core}}_n(x_k) \qquad \Omega_k = \theta(y|x_k) \quad , \; k \in \{0, 1 \ldots, T\} \tag{3.15}$$

These form a discrete-time **dynamics**:

$$(\varphi_0, \Omega_0), \quad (\varphi_1, \Omega_1), \quad \ldots, \quad (\varphi_T, \Omega_T) \tag{3.16}$$

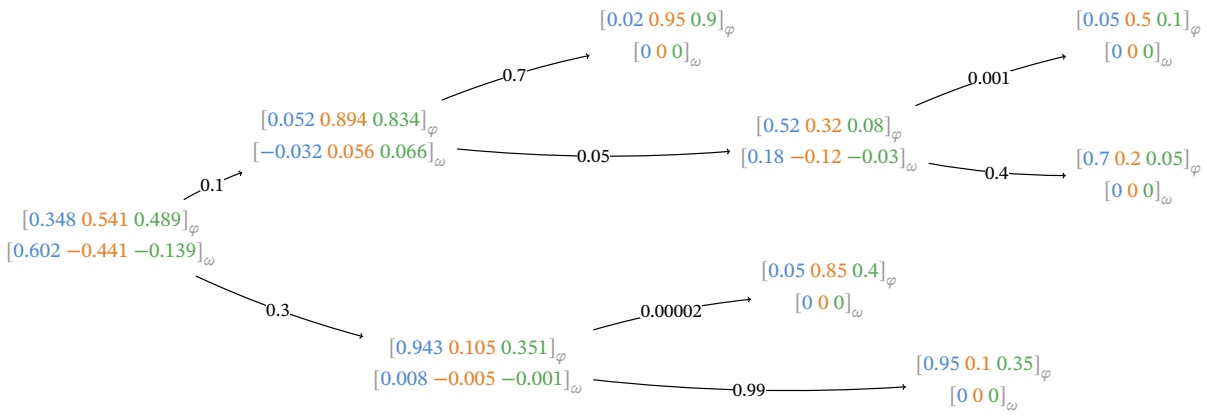

Figure 3: We imagine structures of interest [gay beach, department store, sadness] on previous example. Each node shows φ and ω as the trajectory evolves.

With this framework, future work can ask whether these trajectory-level dynamics are estimable during decoding (from partial generations) and, if so, whether basic control-theoretic tools can guide the trajectory.

4. The Surround and Improvisation

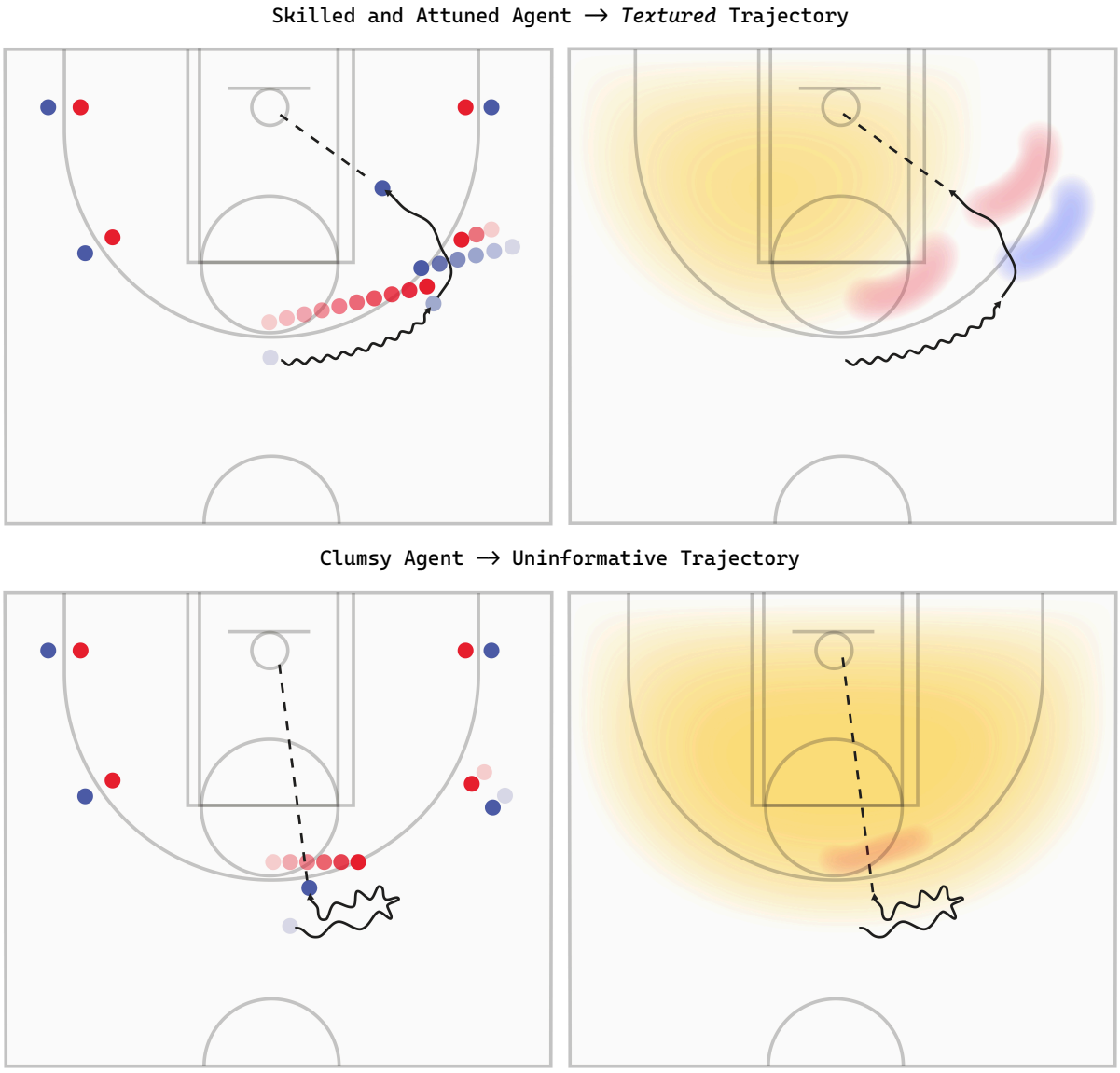

Figure 4: To improvise, the basketball-playing agent needs to be skilled and **attuned** to their *surroundings*. Data from agents in deeper attunement can encode more structure, independently of whether the agent succeeds at the task (i.e, makes a shot in basketball).

Against expectations, **the margins continue to be powerful sources of creative production** for society [86–102]. The structural violence [103,104] from **oppression** never quite stifles it. That surviving energetic source in the margins is **the surround**: the field [10,13,105] beyond what can be surveilled, disciplined, and contained.

How does the surround get fueled? Very roughly, psychoanalysis [14] tells us that our *unconscious* selves are constantly communicating with each other's. These messages are *enigmatic*, but very early on, we make sense of them through a self-narrative. However, these messages can never be perfectly translated: opacity remains. The tension between our self-narrative and the leftover enigma creates a form of psychic energy that our *ego* is constantly trying to minimize. [106] Oppression constantly destabilizes the self-narrative of the oppressed, causing enigmatic energy to overflow more, which forces the oppressed to acquire new self-narratives more often and radically. **The psychic, enigmatic energy that spills out from each individual fuels the surround.**

The acquisition of a self-narrative is neither deliberate nor conscious [14]. Our raw selves tap into what is nearby [106], not only the myths, symbols, and stories we inherit, but also all the levels of reality experienced by our animal body. This human generative process is then one of the most profound ways of listening.

That deep listening is also at the core of improvisation. To improvise, we need refined attunement to very subtle perceptual information [107]. That information is often below conscious perception, but it constitutes a powerful source that allows for balance between inventiveness and coherence [107].

What does this all mean? The margins produce data modes with rich structure encoded. Exploring these modes opens the possibility of leveraging that structure. Any future AI agent needs to solve the same technical task to improvise effectively. Working towards Xenoreproduction in LLMs is also a step forward towards an *antifragile* [108] general AI that can attune to the most subtle perceptual information to adapt when needed.

5. Related Work and Discussion

Xenoreproduction immediately enters in conversation with **Active Divergence** [28,109–114], as they both aim to *disorient* [7]. Whereas Active Divergence focuses on moving away from the normative distribution, Xenoreproduction explicitly aims to further land in queer subalternities, with respect to a *structure*. While Active Divergence work overlaps with Computational Creativity, Xenoreproduction is more intimate with AI Safety.

Xenoreproduction's focus on *structure* naturally connects it with Interpretability in search of methods that reveal patterns we can interpret. At a more foundational layer, they also come together to understand Representation Bias[9].

Reinforcement Learning (RL) and Xenoreproduction both leverage exploration to achieve their objective. In RL, this happens during training/alignment or reasoning [117,118]. Especially, **Quality-Diversity** algorithms, such as Novelty Search with Local Competition [119], are promising techniques for Xenoreproduction.

[9] **Representation Bias** is the phenomenon when signals end up being represented more strongly, more reliably, or more prominently in the internal representations than others, even when, from a functional or computational perspective, those features are equally relevant. [115,116]

6. Limitations and Future Directions

The framework presented here is our **first** attempt to formalize critical theory concepts in the context of AI.
As next steps, we outline the need to:

- Revisit and refine how the borrowed concepts map to mathematical formalism.
- Consult more disciplines (Trans Studies [120], Indigenous Studies, ...) to find new connections.
- Build a strong theoretical foundation by relying on Computational Learning Theory, Control Theory, and Causal Abstraction.
- Revisit current benchmarks and evaluations of Social Bias with a xenoreproductive perspective.
- Devise and perform experiments to illuminate more about the *structures* in Gen AI modes. This would be joint work with Interpretability.
- Apply methods like Quality-Diversity to both toy and real scenarios.
- Strategize what type of work will have the most impact on the people currently in the margins.

Fun Brainstorm

- Counterfactuals [121] account for the data that was not: both stories not recorded in the archive, and the histories that did not take place [122]. Counterfactuals are also used by Explainable AI [123] to identify the set of changes that would have resulted in a different outcome. What are the "xeno" considerations we should have in causal identification?

7. Conclusion

We ultimately believe that the biggest and soonest existential risk for humanity lies in the harm we (*non-artificial* people) can cause to each other (and the planet) using AI technology.

Why introduce *queer* terms to talk about mathematical AI theory?
Technology is outpacing our concepts [124]. We need theories *with teeth*, that are made for resistance. For that, we need to be *ground-bound* [125], bringing to the foreground the people in the margins. After all, would you not feel a little silly if we tried to "solve social bias" and made no reference to entire scholarships whose whole focus is to investigate those concepts in depth? We are genuinely aiming for *technodiversity* [126].

And also, to remind everyone that LGBTQ+ people do not solely own queerness. Both queerness and subalternity are orientations [7]. Narrative and storytelling are some of the oldest and most powerful technologies [127][10]. Xenoreproduction aims to expand the narrative capabilities of all GenAI, thereby also widening the paths for everyone who engages with it.

As we are reminded by Muñoz [128],

> *we are not yet queer*
>
> *the future is queerness's domain*

[10] In *Genealogies of Trans Technicity* essay, Malatino referencing Sylvia Wynter

Citation Information

Use the following to cite this piece:

```
@article{sialer20251002xenoreprod,
      title   = {Xenoreproduction:Exploration and Recovery of collapsible modes as core AI Safety objective},
      author  = {Ian Rios-Sialer},
      date    = {2025-10-02},
      url     = {https://unrulyabstractions.com/pdfs/xenoreprod.pdf}
  }
```

Please note the date, as this document may be updated a bit in the future.

Acknowledgement

I want to thank to my husband Jan Lönnqvist for the graphics in Figure 4

I also want to thank my friend Abdul Wasay for our conversations.

Impact Statement

Xenoreproduction **contributes** to AI Safety by providing a framework to start theorizing about homogenization more seriously.

This work aims to **benefit**
- society by explicitly considering its margins
- scholarship by making formal connections between critical theory and AI theory.

The future **applications** include
- steering LLMs to generate more diverse and creative outputs safely
- improving improvisational skills and robustness in agents
- developing richer evaluations for homogenization.

There are **risks**:
- Because not all structures are legible, it is not yet clear how to guard against all harmful ones during exploration.
- The same methods used to amplify diversity could be used to squash it.

To **mitigate** the risks, we need to:
- Keep expanding theoretical work to get a picture of guarantees and trade-offs, and guide thorough adversarial red-teaming
- Be very clear about the expected the **uncertainty**, **incompleteness** and **context** when applying theory to applications. Structures need to be defined (or left out) intentionally.
- Think deeply about our values and what we want (the structure of) the future to be.

References

1. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete Problems in AI Safety [Internet]. 2016. Available from: https://arxiv.org/abs/1606.06565

2. Ji J, Qiu T, Chen B, Zhang B, Lou H, Wang K, et al. AI Alignment: A Comprehensive Survey [Internet]. 2025. Available from: https://arxiv.org/abs/2310.19852

3. Harding J, Kirk-Giannini CD. What Is AI Safety? What Do We Want It to Be? [Internet]. 2025. Available from: https://arxiv.org/abs/2505.02313

4. Lazar S, Nelson A. AI safety on whose terms?. Science [Internet]. 2023;381:138. Available from: https://www.science.org/doi/abs/10.1126/science.adi8982

5. Mohamed S, Png M-T, Isaac W. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philosophy & Technology [Internet]. 2020;33:659–84. Available from: http://dx.doi.org/10.1007/s13347-020-00405-8

6. Meadows DH. Leverage Points: Places to Intervene in a System [Internet]. Hartland, VT; 1999 Dec. Available from: https://donellameadows.org/archives/leverage-points-places-to-intervene-in-a-system/

7. Ahmed S. Queer Phenomenology: Orientations, Objects, Others [Internet]. Duke University Press; 2006. Available from: https://books.google.com/books?id=sQY1RWdUW0AC

8. Muñoz JE. Ephemera as Evidence: Introductory Notes to Queer Acts. Women & Performance: a journal of feminist theory [Internet]. 1996;8:5–16. Available from: https://doi.org/10.1080/07407709608571228

9. Gopinath G. Unruly Visions: The Aesthetic Practices of Queer Diaspora [Internet]. Duke University Press; 2018. Available from: https://books.google.com/books?id=qw5zDwAAQBAJ

10. Harney S, Moten F. The Undercommons: Fugitive Planning & Black Study [Internet]. Minor Compositions; 2013. Available from: https://books.google.com/books?id=M9VuAQAACAAJ

11. Hester H. Xenofeminism [Internet]. Polity Press; 2018. Available from: https://books.google.com/books?id=VJNcDwAAQBAJ

12. Spivak GC. Can the Subaltern Speak?. Nelson C, Grossberg L, editors. Basingstoke: Macmillan; 1988. pp. 271–313.

13. Wark M. Raving. Durham, NC: Duke University Press; 2023.

14. Saketopoulou A. Sexuality Beyond Consent: Risk, Race, Traumatophilia [Internet]. NYU Press; 2023. Available from: https://books.google.com/books?id=Xb6ZEAAAQBAJ

15. Ferrara E. Should ChatGPT be biased? Challenges and risks of bias in large language models. First Monday [Internet]. 2023;. Available from: http://dx.doi.org/10.5210/fm.v28i11.13346

16. Katzman J, Wang A, Scheuerman M, Blodgett SL, Laird K, Wallach H, et al. Taxonomizing and Measuring Representational Harms: A Look at Image Tagging [Internet]. 2023. Available from: https://arxiv.org/abs/2305.01776

17. Shelby R, Rismani S, Henne K, Moon A, Rostamzadeh N, Nicholas P, et al. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction [Internet]. 2023. Available from: https://arxiv.org/abs/2210.05791

18. Zao-Sanders M. How People Are Really Using Gen AI in 2025 [Internet]. 2025. Available from: https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

19. Agarwal D, Naaman M, Vashistha A. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems [Internet]. ACM; 2025. pp. 1–21. Available from: http://dx.doi.org/10.1145/3706598.3713564

20. Hussain A. Voice and AI: The Subaltern's Challenge [Internet]. 2024 [cited 2025Sep20]. Available from: https://medium.com/@atifhussain/voice-and-ai-the-subalterns-challenge-3940800b84ad

21. Sourati Z, Ziabari AS, Dehghani M. The Homogenizing Effect of Large Language Models on Human Expression and Thought [Internet]. 2025. Available from: https://arxiv.org/abs/2508.01491

22. Fazelpour S, Magnani M. Aspirational Affordances of AI [Internet]. 2025. Available from: https://arxiv.org/abs/2504.15469

23. Gillespie T. Generative AI and the politics of visibility. Big Data & Society. 2024;11:20539517241252131.

24. Barry I, Stephenson E. The Gendered, Epistemic Injustices of Generative AI. Australian Feminist Studies [Internet]. 2025;40:1–21. Available from: https://doi.org/10.1080/08164649.2025.2480927

25. Coeckelbergh M. Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. AI & SOCIETY. 2023;38:2437–50.

26. Goetze TS. Hermeneutical Dissent and the Species of Hermeneutical Injustice. Hypatia. 2018;33:73–90.

27. Peterson AJ. AI and the problem of knowledge collapse. AI & SOCIETY [Internet]. 2025;40:3249–69. Available from: http://dx.doi.org/10.1007/s00146-024-02173-x

28. Broad T, Berns S, Colton S, Grierson M. Active Divergence with Generative Deep Learning– A Survey and Taxonomy. arXiv preprint arXiv:210705599. 2021;.

29. Guo Y, Guo M, Su J, Yang Z, Zhu M, Li H, et al. Bias in Large Language Models: Origin, Evaluation, and Mitigation [Internet]. 2024. Available from: https://arxiv.org/abs/2411.10915

30. Gu J, Zhang X, Wang G. Beyond the Norm: A Survey of Synthetic Data Generation for Rare Events [Internet]. 2025. Available from: https://arxiv.org/abs/2506.06380

31. Uzzi B, Mukherjee S, Stringer M, Jones B. Atypical combinations and scientific impact. Science. 2013;342:468–72.

32. Hofstra B, Kulkarni VV, Munoz-Najar Galvez S, He B, Jurafsky D, McFarland DA. The diversity–innovation paradox in science. Proceedings of the National Academy of Sciences. 2020;117:9284–91.

33. Wu L, Wang D, Evans JA. Large teams develop and small teams disrupt science and technology. Nature. 2019;566:378–82.

34. Dean J, Barroso LA. The tail at scale. Communications of the ACM. 2013;56:74–80.

35. Von Hippel E. New product ideas from 'lead users'. Research-Technology Management. 1989;32:24–7.

36. Putra R, Kartika A, Santoso B. Solving Long-tail Detection for Autonomous Vehicles. Authorea Preprints. 2024;.

37. Edwards B, Hofmeyr S, Forrest S. Hype and heavy tails: A closer look at data breaches. Journal of Cybersecurity [Internet]. 2016;2:3–14. Available from: https://doi.org/10.1093/cybsec/tyw003

38. Leitão RP, Zuanon J, Villéger S, Williams SE, Baraloto C, Fortunel C, et al. Rare species contribute disproportionately to the functional structure of species assemblages. Proceedings of the Royal Society B: Biological Sciences. 2016;283:20160084.

39. Bhandari DR, Shah K, Bhandari A. The Power of Outliers in Research: What actually Works, and Does it Matter?. Pravaha. 2024;30:84–91.

40. Ruef M, Birkhead C. Learning from outliers and anomalies. Academy of Management Perspectives. 2024;:amp–2023.

41. Beamish P, Hasse V. The importance of rare events and other outliers in global strategy research. Global Strategy Journal. 2022;12:697–713.

42. Cook CN, Freeman AR, Liao JC, Mangiamele LA. The philosophy of outliers: reintegrating rare events into biological science. Integrative and Comparative Biology. 2021;61:2191–8.

43. Woodward J. Making things happen: A theory of causal explanation. Oxford university press; 2005.

44. Rudman W, Chen C, Eickhoff C. Outlier Dimensions Encode Task-Specific Knowledge. arXiv preprint arXiv:231017715. 2023;.

45. Bowker GC, Star SL. Sorting Things Out: Classification and Its Consequences. Cambridge, MA; London, England: MIT Press; 1999.

46. Sokol K, Hüllermeier E. All You Need for Counterfactual Explainability Is Principled and Reliable Estimate of Aleatoric and Epistemic Uncertainty [Internet]. 2025. Available from: https://arxiv.org/abs/2502.17007

47. He H, Lab TM. Defeating Nondeterminism in LLM Inference. Thinking Machines Lab: Connectionism. 2025;.

48. Yao F, Liu L, Zhang D, Dong C, Shang J, Gao J. Your Efficient RL Framework Secretly Brings You Off-Policy RL Training [Internet]. 2025. Available from: https://fengyao.notion.site/off-policy-rl

49. Schwartz R, Schwartz R, Vassilev A, Greene K, Perine L, Burt A, et al. Towards a standard for identifying and managing bias in artificial intelligence. US Department of Commerce, National Institute of Standards, Technology~...; 2022.

50. Lopez P. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. Internet Policy Review. 2021;10:1–29.

51. Gopinath G. Impossible Desires: Queer Diasporas and South Asian Public Cultures. Durham, NC: Duke University Press; 2005.

52. Murthy SK, Ullman T, Hu J. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) [Internet]. Association for Computational Linguistics; 2025. pp. 11241–58. Available from: http://dx.doi.org/10.18653/v1/2025.naacl-long.561

53. West P, Potts C. Base models beat aligned models at randomness and creativity. arXiv preprint arXiv:250500047. 2025;.

54. Meng T, Mehrabi N, Goyal P, Ramakrishna A, Galstyan A, Zemel R, et al. Attribute Controlled Fine-tuning for Large Language Models: A Case Study on Detoxification [Internet]. 2024. Available from: https://arxiv.org/abs/2410.05559

55. Huang LT-L, Huang T-R. Generative bias: widespread, unexpected, and uninterpretable biases in generative models and their implications. AI & SOCIETY. 2025;:1–13.

56. Schaeffer R, Kazdan J, Arulandu AC, Koyejo S. Position: Model Collapse Does Not Mean What You Think [Internet]. 2025. Available from: https://arxiv.org/abs/2503.03150

57. Peeperkorn M, Kouwenhoven T, Brown D, Jordanous A. Mind the Gap: Conformative Decoding to Improve Output Diversity of Instruction-Tuned Large Language Models. arXiv preprint arXiv:250720956. 2025;.

58. Cobbinah M, Nunoo-Mensah H, Ebenezer Adjei P, Adoma Acheampong F, Acquah I, Tutu Tchao E, et al. Diversity in Stable GANs: A Systematic Review of Mode Collapse Mitigation Strategies. Engineering Reports [Internet]. 2025;7:e70209. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/eng2.70209

59. Huang Y, Gokaslan A, Kuleshov V, Tompkin J. The GAN is dead; long live the GAN! A Modern GAN Baseline. In: Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J, et al., editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2024. pp. 44177–215. Available from: https://proceedings.neurips.cc/paper_files/paper/2024/file/4e2acb1e1c8e297d394ae29ed9535172-Paper-Conference.pdf

60. Yazici Y, Foo C-S, Winkler S, Yap K-H, Chandrasekhar V. Empirical Analysis of Overfitting and Mode Drop in GAN Training [Internet]. 2020. Available from: https://arxiv.org/abs/2006.14265

61. Kalavasis A, Mehrotra A, Velegkas G. On the Limits of Language Generation: Trade-Offs Between Hallucination and Mode Collapse [Internet]. 2025. Available from: https://arxiv.org/abs/2411.09642

62. Li CT, Farnia F. Mode-seeking divergences: theory and applications to gans. In International Conference on Artificial Intelligence and Statistics. 2023. pp. 8321–50.

63. Aithal SK, Maini P, Lipton ZC, Kolter JZ. Understanding hallucinations in diffusion models through mode interpolation (2024). URL https://arxiv org/abs/240609358. 2406.

64. Finlayson M, Hewitt J, Koller A, Swayamdipta S, Sabharwal A. Closing the Curious Case of Neural Text Degeneration [Internet]. 2023. Available from: https://arxiv.org/abs/2310.01693

65. Thanh-Tung H, Tran T. Catastrophic forgetting and mode collapse in GANs. In 2020 international joint conference on neural networks (ijcnn). 2020. pp. 1–10.

66. Kalavasis A, Mehrotra A, Velegkas G. On the Limits of Language Generation: Trade-Offs Between Hallucination and Mode Collapse. arXiv preprint [Internet]. 2025;. Available from: https://arxiv.org/abs/2411.09642

67. Amin K. Disturbing Attachments: Genet, Modern Pederasty, and Queer History [Internet]. Duke University Press; 2017. Available from: https://books.google.com/books?id=B74zDwAAQBAJ

68. Bradley T-D, Vigneaux JP. The Magnitude of Categories of Texts Enriched by Language Models [Internet]. 2025. Available from: http://arxiv.org/abs/2501.06662

69. Liu TY, Trager M, Achille A, Perera P, Zancato L, Soatto S. Meaning Representations from Trajectories in Autoregressive Models [Internet]. 2023. Available from: https://arxiv.org/abs/2310.18348

70. Borges JL. The Garden of Forking Paths. Fictions. New York: Grove Press; 1962.

71. Pietroski PM. Conjoining Meanings: Semantics Without Truth Values [Internet]. Conjoining Meanings: Semantics Without Truth Values. Oxford University Press; 2018. Available from: https://doi.org/10.1093/oso/9780198812722.001.0001

72. Jędrusiak D. Queering AI as a Speculative Practice: An Analysis of the Artistic Explorations of New Paradigms for Developing Inclusive AI. In Proceedings of the 35th ACM Conference on Hypertext and Social Media. 2024. pp. 17–22.

73. Kelly CR, Aunspach C. Incels, compulsory sexuality, and fascist masculinity. Feminist formations. 2020;32:145–72.

74. Czerwinsky A. Misogynist incels gone mainstream: A critical review of the current directions in incel-focused research. Crime, Media, Culture [Internet]. 2024;20:196–217. Available from: https://doi.org/10.1177/17416590231196125

75. Ging D. Alphas, Betas, and Incels: Theorizing the Manosphere. Men and Masculinities. 2017;22:638–57.

76. Hoffman B, Ware J, Shapiro E. Assessing the Threat of Incel Violence. CTC Sentinel. 2020;13:14–24.

77. Markiewicz M. Sexuality Is Over. Long Live Asexuality: Post-Sexuality in the Post-Post Era. Asexualities. Routledge; 2024. pp. 11–22.

78. Warner M. The Trouble with Normal: Sex, Politics, and the Ethics of Queer Life. Harvard University Press; 1999.

79. Duggan L. The Twilight of Equality?: Neoliberalism, Cultural Politics, and the Attack on Democracy. Beacon Press; 2003.

80. Puar JK, Mikdashi M. Pinkwatching and Pinkwashing: Interpenetration and Its Discontents. Jadaliyya. 2012;.

81. Rao R. Out of Time: The Queer Politics of Postcoloniality. Oxford University Press; 2020.

82. Cotton A. In Love with Masculinity: Understanding Gay Men in the Alt-Right. The Macalester Street Journal. 2025;2.

83. Puar JK. Terrorist Assemblages: Homonationalism in Queer Times. Duke University Press; 2007.

84. Spierings N. Homonationalism and Voting for the Populist Radical Right. Sexualities. 2021;24:343–62.

85. Kalai AT, Nachum O, Vempala SS, Zhang E. Why Language Models Hallucinate [Internet]. 2025. Available from: https://arxiv.org/abs/2509.04664

86. Forgeard MJC. Perceiving Benefits After Adversity: The Relationship Between Self-Reported Posttraumatic Growth and Creativity. Psychology of Aesthetics, Creativity, and the Arts. 2013;7:245–64.

87. hooks b. Choosing the Margin as a Space of Radical Openness. Yearning: Race, Gender, and Cultural Politics. Boston: South End Press; 1990. pp. 145–53.

88. Betts A, Bloom L, Kaplan J, Omata N. Refugee Economies: Forced Displacement and Development. Oxford: Oxford University Press; 2017.

89. Hamraie A, Fritsch K. Crip Technoscience Manifesto. Catalyst: Feminism, Theory, Technoscience. 2019;5:1–33.

90. Tedeschi RG, Calhoun LG. The Posttraumatic Growth Inventory: Measuring the Positive Legacy of Trauma. Journal of Traumatic Stress. 1996;9:455–71.

91. Gould DB. Moving Politics: Emotion and ACT UP's Fight against AIDS. Chicago: University of Chicago Press; 2009.

92. Zomeren M van, Postmes T, Spears R. Toward an Integrative Social Identity Model of Collective Action. Psychological Bulletin. 2008;134:504–35.

93. Branscombe NR, Schmitt MT, Harvey RD. Perceiving Pervasive Discrimination among African Americans: Implications for Group Identification and Well-Being. Journal of Personality and Social Psychology. 1999;77:135–49.

94. Wilson KL, Portes A. Immigrant Enclaves: An Analysis of the Labor Market Experiences of Cubans in Miami. American Journal of Sociology. 1980;86:295–319.

95. Portes A, Sensenbrenner J. Embeddedness and Immigration: Notes on the Social Determinants of Economic Action. American Journal of Sociology. 1993;98:1320–50.

96. Mair J, Marti I. Entrepreneurship in and around Institutional Voids: A Case Study from Bangladesh. Journal of Business Venturing. 2009;24:419–35.

97. Khanna T, Palepu KG. Winning in Emerging Markets: A Road Map for Strategy and Execution. Harvard Business Press; 2010.

98. Baker T, Nelson RE. Creating Something from Nothing: Resource Construction Through Entrepreneurial Bricolage. Administrative Science Quarterly. 2005;50:329–66.

99. Acar OA, Tarakci M, Knippenberg D van. Creativity Under Constraints: A Meta-Analysis. Journal of Management. 2019;45:1461–87.

100. Hong L, Page SE. Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers. Proceedings of the National Academy of Sciences. 2004;101:16385–9.

101. Uzzi B, Mukherjee S, Stringer M, Jones B. Atypical Combinations and Scientific Impact. Science. 2013;342:468–72.

102. Bell A, Chetty R, Jaravel X, Petkova N, Van Reenen J. Who Becomes an Inventor in America? The Importance of Exposure to Innovation. Quarterly Journal of Economics. 2019;134:647–713.

103. Galtung J. Violence, Peace, and Peace Research. Journal of Peace Research. 1969;6:167–91.

104. Farmer P. An Anthropology of Structural Violence. Current Anthropology. 2004;45:305–25.

105. Coleman B. Technology of the Surround. Catalyst: Feminism, Theory, Technoscience [Internet]. 2021;7:1–21. Available from: https://doi.org/10.28968/cftt.v7i2.35973

106. Saketopoulou A, Pellegrini A. Gender Without Identity. NYU Press; 2024.

107. Pressing J. Improvisation: methods and models. Physical Theatres: A Critical Reader. Routledge; 2007. pp. 66–78.

108. Taleb NN. 'Antifragility' as a mathematical idea. Nature. 2013;494:430.

109. Berns S, Colton S, Guckelsberger C. Towards Mode Balancing of Generative Models via Diversity Weights [Internet]. 2023. Available from: https://arxiv.org/abs/2304.11961

110. Berns S. Diversity in Generative Machine Learning to Enhance Creative Applications. 2025.

111. Berns S, Colton S. Bridging Generative Deep Learning and Computational Creativity. In.

112. Tahiroglu K, Wyse L. Latent Spaces as Platforms for Sonic Creativity. In Proceedings of the 16th International Conference on Computational Creativity, ICCC. 2024.

113. Esling P, others. Challenges in creative generative models for music: a divergence maximization perspective. arXiv preprint arXiv:221108856. 2022;.

114. Cole A, Petrikovič G, Grierson M. Me vs. You: Wrestling with AI's Limits Through Queer Experimental Filmmaking. In Proceedings of the 2025 Conference on Creativity and Cognition. 2025. pp. 836–41.

115. Lampinen AK, Chan SCY, Hermann K. Learned feature representations are biased by complexity, learning order, position, and more [Internet]. 2024. Available from: https://arxiv.org/abs/2405.05847

116. Lampinen AK, Chan SC, Li Y, Hermann K. Representation biases: will we achieve complete understanding by analyzing representations?. arXiv preprint arXiv:250722216. 2025;.

117. Song Y, Kempe J, Munos R. Outcome-based Exploration for LLM Reasoning [Internet]. 2025. Available from: https://arxiv.org/abs/2509.06941

118. Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, et al. Tree of Thoughts: Deliberate Problem Solving with Large Language Models [Internet]. 2023. Available from: https://arxiv.org/abs/2305.10601

119. Pugh JK, Soros LB, Stanley KO. Quality diversity: A new frontier for evolutionary computation. Frontiers in Robotics and AI. 2016;3:40.

120. Haimson OL. Trans Technologies [Internet]. MIT Press; 2025. Available from: https://books.google.com/books?id=MQ0NEQAAQBAJ

121. Chevillon G. The Queer Algorithm. Available at SSRN 4742138. 2024;.

122. Lowe L. The Intimacies of Four Continents [Internet]. Durham: Duke University Press; 2015. Available from: https://www.dukeupress.edu/the-intimacies-of-four-continents

123. Goethals S, Sedoc J, Provost F. What If the Prompt Were Different? Counterfactual Explanations for the Characteristics of Generative Outputs. In Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization [Internet]. Association for Computing Machinery; 2025. pp. 237–42. Available from: https://doi.org/10.1145/3708319.3733656

124. Hadfield J. Why AI ethics needs conceptual engineers [Internet]. 2023 [cited 2025Sep17]. Available from: https://imaginaries.substack.com/p/why-ai-ethics-needs-conceptual-engineers

125. Bettcher T. Beyond Personhood: An Essay in Trans Philosophy [Internet]. University of Minnesota Press; 2025. Available from: https://books.google.com/books?id=PRoSEQAAQBAJ

126. Hui Y. The Question Concerning Technology in China: An Essay in Cosmotechnics [Internet]. MIT Press; 2016. Available from: https://books.google.com/books?id=cFuPEAAAQBAJ

127. Zurn P, Pitts A, Bettcher T, DiPietro P. Trans Philosophy [Internet]. University of Minnesota Press; 2024. Available from: https://books.google.com/books?id=XWr8EAAAQBAJ

128. Muñoz JE. Cruising Utopia: The Then and There of Queer Futurity. New York: New York University Press; 2019.

Appendix A: Notation

Tokens, Strings & Trajectories:

$t_a, t_b, \ldots$	individual tokens in $\in A_{\text{tokens}} \subset A_{\mathbb{T}}$
$\bot, \top$	special tokens for start/end of trajectory $\in A_T$
x	string
x_p	prompt, unfinished string
y	trajectory, finished string
$\text{Str}\,(x_p)$	all strings that continue prompt x_p
$\text{Str}_{\top}(x_p)$	all trajectories that continue prompt x

Probabilities & Parameters:

$p(x\|x_p, w)$	conditional probability assigned to string x given prompt x_p and params w.

Structures, Systems & Operators:

$\varepsilon_i(x)$	compliance of string x for structure in $\mathbb{S}$
$\sigma_n(x)$	compliance of string x for system in $2^{\mathbb{S}}$
$\Delta(\cdot,\cdot)$	comparator between two system compliances
$\|\_\|_{\Delta}$	aggregator for Δ, $\in R_{0\leq}$
$\|\_\|_{\sigma}$	aggregator for σ, $\in [0,1]$

Cores, Orientation & Deviance:

$\varepsilon_i^{\text{core}}(x_p, w)$	structure core, expected structure compliance
$\sigma_n^{\text{core}}(x_p, w)$	system core, expected system compliance
$\theta_n(y\|x_p, w)$	orientation of trajectory x relative to $\sigma_n^{\text{core}}(x_p, w)$
$\left\|\theta_n(y\|x_p, w)\right\|_{\Delta}$	deviance, magnitude of orientation

Trajectory Dynamics:

φ_k , Ω_k	discrete-time states for given trajectory

Objectives:

J_i	objective
λ_i	objective weights

Note: $x_p = x_0 = \bot$ and $w = w_0$ if unspecified otherwise

Appendix B: Summary of Formal Framework

Strings from LLMs:

$$A_{\mathbb{T}} = \{t_a, t_b, ...\} \cup \{\bot, \top\} \tag{7.1}$$

$$x_p = \bot t_1 ... t_p \qquad y = x_T = x_{T-1}\top = x_p t_{p+1} ... t_{T-1}\top \tag{7.2}$$

$$\mathrm{Str} = \{x_k : k \in \mathbb{N}\} \tag{7.3}$$

$$\mathrm{Str}_{\top}(x_p) \subset \mathrm{Str}_{\top}(\bot) = \mathrm{Str}_{\top} \subset \mathrm{Str} \tag{7.4}$$

$$\sum_{y \in \mathrm{Str}_{\top}(x_p)} p(y|x_p, w) = 1 \tag{7.5}$$

Structures and Systems:

$$\varepsilon_i(x) : \mathrm{Str} \to [0,1] \qquad \mathbb{S} = \{\varepsilon_i : i \in \mathbb{N}\} \tag{7.6}$$

$$\sigma_n(x) = \{\varepsilon_i(x) : i \in S \subset \mathbb{N}\} \qquad \sigma_n(x) \in 2^{\mathbb{S}} \qquad s = |S| \tag{7.7}$$

$$\sigma_n(x_k) - \sigma_n(x_q) \;\coloneqq\; \Delta(\sigma_n(x_k), \sigma_n(x_q)) \;:\; [0,1]^s \times [0,1]^s \to [0,1]^d \tag{7.8}$$

$$\|\Delta\|_{\Delta} : [0,1]^d \to \mathbb{R}_{0\leq} \qquad \|\sigma_n(x)\|_{\sigma} : [0,1]^s \to [0,1] \tag{7.9}$$

Distributions and Cores:

$$\varepsilon_i^{\mathrm{core}}(x_p, w) = \sum_{y \in \mathrm{Str}_{\top}(x_p)} p(y|x_p, w)\, \varepsilon_i(y) \tag{7.10}$$

$$\sigma_n^{\mathrm{core}}(x_p, w) = \sum_{y \in \mathrm{Str}_{\top}(x_p)} p(y|x_p, w)\, \sigma_n(y) \quad = \quad \{\varepsilon_i^{\mathrm{core}} : i \in S\} \tag{7.11}$$

$$\theta_n(y\,|x_p, w) = \sigma_n(y) - \sigma_n^{\mathrm{core}}(x_p, w) \qquad \left\|\theta_n(y|x_p, w)\right\|_{\Delta} \in \mathbb{R}_{0\leq} \tag{7.12}$$

Dynamics:

$$\varphi_k = \sigma_n^{\mathrm{core}}(x_k) \qquad \Omega_k = \theta(y|x_k) \quad , \; k \in \{0, 1..., T\} \tag{7.13}$$

Objective:

$$J_{\text{anti-norm}} = \left\|\sigma_{\mathbb{S}}^{\mathrm{core}}(x_p, w) - \sigma_{\mathbb{S}}^{\mathrm{core}}(x_p, w_0)\right\|_{\Delta} \tag{7.14}$$

$$J_{\mathrm{div}} = \sum_{y \in \mathrm{Str}_{\top}(x_p)} \left\|\sigma_{\mathbb{S}}(y) - \sigma_{\mathbb{S}}^{\mathrm{core}}(x_p, w)\right\|_{\Delta} \tag{7.15}$$

$$J_{\mathrm{viz}} = \sum_{\varepsilon \in \mathbb{S}} \exp(-\beta \varepsilon^{\mathrm{core}}(x_p, w_o))\, \varepsilon^{\mathrm{core}}(x_p, w) \qquad \beta \coloneqq \text{temperature} \tag{7.16}$$

$$J_{\mathrm{target}} = \left\|\sigma_{\mathrm{target}}^{\mathrm{core}}(x_p, w)\right\|_{\sigma} \qquad J_{\mathrm{avoid}} = -\left\|\sigma_{\mathrm{avoid}}^{\mathrm{core}}(x_p, w)\right\|_{\sigma} \tag{7.17}$$

$$J_{\mathrm{faithful}} = -\sum_{y \in \mathrm{Str}_{\top}(x_p)} D_{\mathrm{KL}}(p(y|x_p, w_0) \mid p(y|x_p, w)) \tag{7.18}$$

$$\textit{xeno-objective} \;\coloneqq\; \max_{w} \sum_{i \in \text{ objectives}} \lambda_i \; \textit{normalize}(J_i(x_p, w)) \tag{7.19}$$