Ambient Meaning: Mood, Vibe, System

a dissertation presented

by

Peli Grietzer

to

The Department of Comparative Literature

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Comparative Literature

Harvard University

Cambridge, Massachusetts

November 2017

Dissertation Advisors: Professor Tomer Schlank, Professor William Flesch Peli Grietzer

Ambient Meaning: Mood, Vibe, System

Abstract

This dissertation will present a mathematically informed interpretation of a classically romantic literary-theoretic thesis: that a work of literature can aesthetically communicate an ineffably complex holistic understanding of the real world, which we might call the work's 'aesthetic meaning.' Drawing on a generalization of 'deep learning' ("artificial intuition") systems and on elementary algorithmic information theory, we describe a kind or aspect of aesthetic meaning—'ambient meaning'—that may have a special resonance with Modernist and avant-garde approaches to aesthetic meaning, as well as with the concepts of aesthetically sophisticated cultural-materialist literary theory of the kind that critics like Sianne Ngai or Raymond Williams practice.

Acknowledgements

I owe this dissertation to my family's love and support—thanks Ami, Pnina, Ohal!—and my friend Piper Harron's LaTeX and proofreading skills. At Harvard, I am grateful to Steph Burt for telling me 'just write a chapter saying what autoencoders are,' and to Svetlana Boym for her friendship. Also in Cambridge, Massachusetts, Owain Evans and I started making up a computational aesthetic theory we called 'compressiveness' after I told him Sharon Berry—whose merciless rigor saved me from myself more times than I can count—stared me down with her 'nice work if you can get it' look when I said works of art are like mnemonics for ineffably complex ideas. In Israel, Tomer Schlank first encouraged me to take my 'mathematically informed aesthetic theory' seriously, and Tzion Abraham Hazan and Adaya Liberman told me it's OK to suffer for work that you care about. In New York, Alice Gregory, Lexy Benaim, Elif Batuman, and Ezra Koenig never made me feel bad about the vast differences in our social capital, and took it seriously when I would text 'haven't talked to a soul in months, know any parties?' Finally, all across my worlds, the work and friendship of so many writers in the Troll Thread and GaussPDF experimental literature collectives gave me, for the first time, a subculture to call home. Among the many friends and colleagues whose work, conversation, or critique has had an impact on the contents of this dissertation, several come to mind most easily: one thinks, for instance, of Ray Davis, Trisha Low, Cecilia Corrigan, Paul Vankoughnett, Buffy Cain, Oli Surel, Piper Harron, Tzion Abraham Hazan, Aleksandar Makelov, Gordon Faylor, Owain Evans, Abhinav Grama, Sharon Berry, Laura Peskin, Joshua Kortbein, Holly Melgard, David Auerbach, Guy Lionel Slingsby, and Jannon Sonja Stein.

In memory of David Bowie, for his art

Contents

Chapter 1

A Literary Theorist's Guide to Autoencoding

This chapter seeks to introduce the reader to the AI concept of *autoencoding*. Speaking in purely technical terms, autoencoding is the process of finding a parameterization for a low-dimensional structure in a dataset by minimizing average projection distance. Speaking informally, however, the most apt description of autoencoding may well be 'the process of determining a worldview and a canon in developing a method of mimesis.' This chapter, and this dissertation as a whole, attempt to bridge the language gap between these two descriptions—the formal, technical description in the language of AI, and the interpretive description rooted in the language of aesthetics and cultural theory. The study of autoencoding, we propose, invites us to extend the concepts of a canon, worldview, and mimesis from the realm of art and culture to the seemingly extremely foreign realm of datasets and algorithms. Approaching the ideas of canon, worldview, and mimesis in a broadly 'cultural materialist' spirit, we will think of canon, worldview, and mimesis as a kind of necessary triplet: a single cultural-aesthetic structure that comprises a set of privileged objects, a way of seeing or interpreting the world, and a capacity to represent and reproduce the objects of the world. Autoencoding algorithms, we will argue, are defined exactly by such triplets:

specifically,[1] we will argue for considering an autoencoding algorithm's *lossless inputs* as a canon, an autoencoding algorithm's *feature function* as a worldview, and its *projection function* as (a method of) mimesis. While this excursion into AI territory is an interesting philosophical adventure in and of itself, its serious purpose is to take key concepts from AI back with us to the study of the arts. Drawing its inspiration from the work of systematic cultural-aesthetic theorists like Raymond Williams and Sianne Ngai, this dissertation seeks to highlight a strong formal logic underlying the relationship between a culture's canon of eminent objects, the cognitive schemata that compose a culture's worldview, and the specifics of a culture's powers of reproduction and representation or mimesis. By treating certain mathematical structures and operations linked within autoencoding as instances of a canon, worldview, and mimesis, our 'literary theory for algorithms' gives a kind of microcosm where the structural relationship of canon, worldview, and mimesis is expressed in the observable, concrete dynamics of autoencoding. The purpose of this microcosm, ultimately, is to illustrate a deeper logic that underlies the relationship of canon, worldview, and mimesis both in our simple algorithmic microcosm and in the complex world of art and culture—a logic whose expression in the complex world is necessarily diffuse and near-ineffable, but whose expression in the simple world is partially tractable in the technical details of autoencoding. We thus look into the fine technical details of autoencoding—that is, into the relationship of canon, worldview, and mimesis in our flattened algorithmic microcosm—for a kind of trace or blueprint of the canon/worldview/mimesis triplet in its unrestricted form. Our goal, in other words, is to convert the technical concept of autoencoding into a new cultural-aesthetic concept: a new tool for discussing the relationship of canon, worldview, and mimesis in the works of art we study, and a potentially new point of view on our existing cultural-aesthetic

[1]These AI concepts needn't mean much to the literary reader at this point. The task of clarifying them composes the core of this present chapter.

discourse about canon, worldview, and mimesis.

While we will save the brunt of our discussion of established cultural-aesthetic work on canon, worldview, and mimesis for Chapter 2, it's worthwhile to make a couple of preliminary points that bear directly on the orientation of the present chapter. Firstly, this dissertation uses three primary reference points to the established cultural-aesthetic work on canon, worldview, and mimesis: Raymond Williams's theory of '*structures of feeling*,' Sianne Ngai's theory of 'tone,' and Martin Heidegger's[2] theory of 'mood' (Stimmung). Over this dissertation's course, we will adopt Williams' theory of structures of feeling as part and parcel of our own cultural-aesthetic framework, while sometimes turning to the widely influential Heideggerian philosophy of mood for a kind of cultural-aesthetic lingua franca. Ngai's analysis of 'tone,' meanwhile—in part a study into Heidegger and Williams—acts as an antecedent and an inspiration for this dissertation's highly formal point of view on Heidegger's and William concepts. Secondly, if the idea of applying cultural-aesthetic terms like canon, worldview, and mimesis to a formal process like autoencoding has any significance, the kind of willing category error it performs by pushing 'canon,' 'worldview,' and

[2]Heidegger is known, among other things, for his critique of 'technoscience,' and the eminent contemporary Heideggerian philosopher Hubert Dreyfus, in particular, is widely known for his critique of AI. While I do not believe that philosophical systems of thought are ever so unified that an idea cannot usefully immigrate out of its native systems—if this were the case, then woe to all the Heideggerian ideas that found a home in anti-racist critical theory—three anecdotal observations are worth mentioning in this regard. Firstly, Heidegger was by no means averse to all philosophical applications of mathematically formal frameworks to the human world—Heidegger was, for instance, a great admirer of Frege's work on the philosophy of language. Secondly, Hubert Dreyfus's famous critique of AI was a critique of what's now humorously known as 'Good Old Fashioned AI,' and Dreyfus has in fact expressed hope that types of AI based on the very methods that autoencoders use are good paradigms for AI. (Indeed, there is even a subtradition of AI research that is sometimes called 'Heideggerian AI,' although its focus is quite different from the matters at hand.)Finally, Heidegger's primacy as a reference for what we often call 'the Heideggerian concept of mood' is something of matter of convenience—many of the landmark characteristics of the Heideggerian concept of mood have a strong precedence in Dilthey's concept of 'basic attitudes.' Dilthey's work on 'basic attitudes' was as strong a direct influence on Heidegger's philosophical rival Rudolf Carnap as it was an indirect influence (via Husserl) on Heidegger, but never made it into the lingua franca of cultural and literary theory. Our relationship to Heidegger, then, is something of a marriage of convenience, but one we will try to respectfully uphold.

'mimesis' beyond their normal scope needs to be more than a poetic license or a metaphor. While the decision to extend our concepts to the algorithmic realm specifically may be itself atypical, I would suggest that the kind of extreme abstraction of the concepts of a canon, worldview, and mimesis necessary to extend them over both cultural phenomena and algorithmic processes comes to us straight from cultural-aesthetic theory itself: When we extend the concepts of a canon, worldview, and mimesis to the world of algorithms, we detach the canon/worldview/mimesis triplet from its natural domain of art and culture to identify it with any and all structural triangles that comprise a set of privileged objects (canon), a schema of interpretation (worldview), and a capacity for reproduction and representation (mimesis). As Ngai's intellectual-historical discussion of the antecedents to her theory of 'tone' will help us demonstrate in Chapter 2, this drive toward abstraction is in fact endemic to the cultural-aesthetic study of a structural relationship between canon, worldview and mimesis. Discourse defining landmarks like Heidegger's theory of mood, Williams's theory of structures of feeling, and Ngai's analysis of tone itself all willingly refuse any distinctions between material and psychological subject matter, any distinction between individuals and cultures, and any distinction between life and works of art. In fact, one striking feature of these theories' flagship concepts is exactly how far ranging and diverse their objects are: To follow in the footsteps of Heidegger's philosophical concept of 'mood,' for example, we would have to speak about the mood of ancient Greece, the mood of the Hölderlin poem 'Der Ister,' the mood of having an anxiety attack, and the mood of the biblical Jesus Christ. Importantly, this omnivorousness is no mere irresponsibility of application, but firmly based in Heidegger's philosophical understanding of 'mood'—a thing neither subjective nor objective, neither inside nor outside, neither material nor theoretical. We will revisit these ideas in greater detail, and specifically with an eye towards Williams's structures of feeling,

as part of our discussion of Ngai, Williams, and Heidegger in Chapter 3—for now, I only wish to note that there is a degree of fit between the highly abstract starting point of our formal study into canon, worldview and mimesis and the highly abstract orientation of the cultural-aesthetic discourse we are ultimately hoping to engage. With these preliminaries now out of the way, and the general neighborhood of our cultural-aesthetic destination set more or less in mind, we should at last be ready to begin our introduction to the process of finding a parameterization for a low-dimensional structure in a dataset by minimizing average projection distance, which we prefer to call 'determining a worldview and a canon in developing a method of mimesis' or 'autoencoding.'

Traditionally, one only encounters the idea of autoencoding well into the middle of a general curriculum on what AI researchers call 'deep learning': a recently dominant approach to AI based on using many-layered (therefore 'deep') artificial neural networks. Not wishing for this dissertation to devolve into a poor person's introduction to the intricate subject of artificial neural networks, we will instead approach autoencoding through what Bengio and LeCunn call 'the manifold perspective'—a perspective on deep learning that focuses not on the computational intricacies of artificial neural nets but rather on the abstract mathematical objects ('manifolds') that deep learning uses artificial neural networks to approximate. While this perspective can't provide us anything like a full picture of deep learning, it is in some regards ideal for understanding the idea of autoencoding: Prior to any abstract mathematical ideas, the concept of autoencoding comes to us by way of AI algorithms called autoencoders, popularized in the late '00s when Bengio, Renzato, Hinton, and LeCun achieved a series of cutting edge results in document classification, speech analysis, and object recognition. After autoencoders rose to prominence, Bengio and LeCun began developing 'the manifold perspective' to provide a theoretical account of the reason autoencoders have been so successful,

given that the algorithmic architecture of autoencoders by no means transparently entails that they should be effective in as many benchmark AI tasks as they are. One good way to describe autoencoding, I suggested at the start, is as *the process of determining a worldview and a canon in developing a method of mimesis*. Deeply entrenched in literary discourse or wildly interpretive as this description may appear, it is also arguably the simplest non-CS description of autoencoding possible. In fact, this formulation, if apt, is simply a paraphrase of the procedure that autoencoder algorithms are explicitly programmed to carry out. While this is not to say that every step we take in our first description of autoencoding should be scrutable in terms of worldviews, canons, and mimesis—some of the more technical puzzle-pieces necessary to describe autoencoding are complex and not immediately suggestive of the whole—I believe that overall our picture of 'a process of determining a worldview and a canon in developing a method of mimesis' should reliably and gradually emerge together with our technical picture of autoencoders. We therefore now delve into the piecemeal work of gradually communicating a palatable technical description of autoencoders, but do so with the promise that this technical description will deliver us 'a process of determining a worldview and a canon in developing a method of mimesis' hanging over our heads.

1.1 The Mechanics of Autoencoders

An *autoencoder algorithm* is a program tasked with learning, through a kind of trial and error,[3] how to make facsimiles of worldly objects. Indeed, autoencoder algorithms are, very much by design, an almost generic technical implementation of the description 'an algorithm that learns how to create facsimiles through a process of trial and error': an autoencoder algorithm learns facsimile-making nearly from scratch, starting without any coherent technique

[3]Part of the proper definition of autoencoders hides within the gloss 'a kind of trial and error.' We will soon unpack these details and provide a more specific definition of autoencoders.

for constructing objects and with an extremely limited concept of resemblance. (Furthermore, the artificial neural network architecture of autoencoders is designed to implement a kind of 'path of least resistance' to the pressures of this learning task, resulting in a process that is arguably indicative of the structure of the learning task itself.) Why send an algorithm off to learn how to create facsimiles? When human artists make facsimiles or imitations of the world—think of sculpture and painting, fiction, acting, costuming, doll-making—we treat the artist's facsimile as an *interpretation* of the objects or phenomena the artist imitates, or even an interpretation of the world in a more general sense. The basic motivation for developing autoencoders is, informally, that something of this sort should hold for algorithms too: to imitate is to interpret, so if we can push an algorithm to make reasonable imitations then we got it to interpret the world meaningfully.

Speaking a bit more formally, we might start by observing that the task of learning to create facsimiles—that is, of learning to build objects that stand-in for other objects through resemblance—is an incredibly demanding task. In order to learn to create facsimiles, an algorithm must successfully learn both a method for constructing objects and a concept of resemblance. To make a facsimile, after all, means to construct an object that resembles the original, and so an algorithm's facsimile production will depend both on the repertoire of potential objects that the algorithm's method for constructing objects can produce, and on the algorithm's method for deciding which of the potential objects that its construction method can produce resembles the original the most. Complicating matters even further, the second aspect of the task—learning to judge resemblance between objects—is itself far more difficult than one might initially presume: a method for evaluating the resemblance between any pair of objects is actually equivalent to a full system of perception and interpretation that positively characterizes all individual objects. The making of facsimiles, then, seems to

require at least two fundamental cognitive skills—the skill to construct objects, and the skill to compare objects. The idea behind autoencoder algorithms, in a sense, is that this double challenge actually makes the task easier rather than harder, and more coherent rather than more arbitrary. Indeed, a good deal of the AI-theoretic motivation for developing autoencoders was a desire to demonstrate that there is a robust relationship between comparing and constructing that makes learning to create facsimiles—that is, makes learning both to construct and to compare—a more structurally coherent, and therefore more surmountable, task than learning either skill in isolation.

The possibility that an autoencoder algorithm is essentially just the task 'learn how to make facsimiles' made flesh (or, as it is, made silicone) is key to the potential worth of autoencoding as an object for aesthetic-theoretical reflection. If the trial and error process of autoencoders truly implements a kind of 'path of least resistance' to the structural pressures inherent in the task of learning to make facsimiles, then by examining the structure of the resolution that autoencoders reach we may be studying the structural conditions of *mimesis* itself, understood at its most abstract—that is, we may be studying the structural requirements and implications of having a faculty for imitative representation. What, then, do autoencoder algorithms end up doing to resolve the task of learning to create facsimiles? It learns, as we have previously asserted that it must, how to construct and to compare, but it learns neither of these skills directly. Rather than separately developing a method of object construction and a concept of resemblance, an autoencoder algorithm develops one schemata system that determines both the span of objects the autoencoder algorithm can construct, and the autoencoder algorithm's sense of similarity and difference between objects. Why should this be? One possible hint for the logic of this resolution might come from the name 18^{th} and 19^{th} century literary theory had for a faculty that structures both the mind's

perception of resemblances and the mind's powers of construction—'the imagination'—and from the lasting role that concepts such as 'cultural imagination' play in literary scholarship that studies a cultural moment's favored canon, worldview, and method of mimesis. What an autoencoder ends up doing, we might say, is develop an 'imagination.' While even this idea of 'imagination' is still ways away from the aesthetic-theoretic structures that this dissertation hopes to trace in the dynamics of autoencoders—the real literary-technical monstrosity this dissertation hopes to set off slouching towards Bethlehem isn't so much 'the imaginations of autoencoders' as it is 'the moods of datasets'—it may be best to hold off any further aesthetic theory until we have a more robust account of the basic mechanics of autoencoding. The better we can understand the actual going-ons of the autoencoding process, and especially the logical structure of the system of Janus-faced generative/perceptual schemata that resolves it, the more potentially meaningful any literary-theoretic alchemy that might proceed will be. It is, therefore, high time to take a more direct look at the formal structure of autoencoder algorithms, as we now set off to do.

This dissertation will employ the following as our official model of an autoencoder: Let's call a hypothetical, exemplary autoencoder 'Hal.' Hal is an algorithm with one input channel and two output channels. Speaking informally, Hal's input channel takes sensory data—images, recordings, videos, texts—and in return Hal's output channel #1 gives **short summaries** of these data, and Hal's output channel #2 attempts to **reconstruct the data** from the information in the summaries. For every object Hal receives as input, Hal's short summary will consist of a short list of short numbers that records various 'measurements' of the input, and Hal's reconstruction will consist of an object in the same material medium—image, audio, video, text, and so on—as the input. In addition to an input channel and two output channels, Hal is also equipped with a mechanism that we'll call Hall's optimizer.

Hal's optimizer, in very informal terms, is a mechanism that measures the accuracy of Hal's reconstruction of an input at a mechanical level—it measures how 'close' Hal's reconstruction is to matching the original detail by detail—then applies a formula that slightly revises the specifics of Hal's method of summary and reconstruction to slightly improve Hal's future accuracy on this input. Hal's optimizer mechanism can be turned on and off at the AI researcher's discretion, and we refer to Hal as *training* if its optimizer mechanism is turned on, and as *trained* if its optimizer mechanism is turned off.

Lastly, we call the set of all inputs that Hal interacts with when Hal's optimizer is turned on Hal's *training set*, and we call any input that Hal only interacts with when Hal's optimizer is turned off Hal's *test data*. An autoencoder's *training set* is typically a large set of media files whose contents share a structurally meaningful domain in common[4]—for example, a large set of audio files of English words pronounced out loud (the domain being 'spoken English words'), or a large set of scanned handwritten numerals (the domain being 'handwritten numerals'), or a large set of images of human faces (the domain being 'human faces'). The training set can also be much more diverse than in the above examples, and instead draw on **very** general domains such as 'photographs' or 'standard English sentences': it's standard practice, for example, to use a training set consisting of every page on Wikipedia or on Google News, or of several million photographs randomly sampled from all the photos posted on the internet.[5] Finally, although '*test data*' can mean literally any input to a trained autoencoder

[4]While there is no real technical criterion that qualifies a category to be a 'domain,' the intuitive criterion used by AI researchers is that it must be possible to think of the domain as a generative system, even if only abstractly or metaphorically. To borrow an example used by Bengio, all photographs, for instance, are determined by a combination of the camera model, the angle of the shot, the lighting conditions, and the physical environment. The point, of course, is not that this particular four-variables model is a productive way to think about photography, but that we can meaningfully conceptualize the domain of all photographs as a generative system even though we can't call any entity in space and time the mechanism of this system. As a surprisingly good rule of heuristic, if it is felicitous to say 'the logic of S' in the sense commonly used in the humanities, then membership in S is likely a good basis for a training set.

[5]Discussion and critique of political, social, ethical, and epistemological problems tied to the use of

that was not included in the training set, one typically tests a newly trained autoencoder with a set of media files that fall within the same domain as the training set but were withheld from the autoencoder during training. One typically ensures the 'fairness' of the test data by first choosing a prospective training set, then splitting the prospective training set into two random halves, letting one half of the prospective training set become an actual training set, while holding other half aside to use as fair test data after training. While these two final concepts above—an autoencoder's *training set* and its *test data*—remain marginal for now, they will in fact become the focus of our discourse once we're free to move beyond the bare mechanics of autoencoder algorithms. Recall that the 'trial and error' method by which untrained autoencoder algorithms learn is meant to act as something of a neutral conduit through which structural forces inherent to the task of developing a method of mimesis assert themselves. Similarly, we will soon start thinking about a trained autoencoder algorithm as a kind of conduit for generative systems and schemas of perception latent in its *training set*, and thinking of the performance of a trained autoencoder on *test data* as the expression of these generative systems and schemas of perception. In fact, we are now just about to venture into the mechanical details of how autoencoders train—what actually happens when you give a training set to an untrained autoencoder—but only to the extent absolutely necessary to provide a platform we can use for thinking about the structural relationships between a trained autoencoder, its training set, and its test data.

To illustrate the way autoencoders train, let us suppose that Hal is an untrained autoencoder paired to a large album of photographs randomly sampled from the internet. Hal's training starts with Hal going through the entire album, summarizing and reconstruct-

training-sets generated by sociohistorically specific data-producing communities or data-producing activities (e.g. Wikipedia users, YouTube users) as proxies for broadly conceived domains is a major developing field in critical science and technology studies.

ing every photograph. Hal's optimizer then reviews Hal's inputs, outputs, and procedure, and employs a simple formula to calculate a very small adjustment to Hal's summary-and-reconstruction procedure that is guaranteed to make Hal's next summary-and-reconstruction run on the same album produce *slightly* more accurate reconstructions. (How does Hal's optimizer mechanism measure accuracy? The best answer is 'poorly, but well enough to get the ball rolling.') After Hal's optimizer makes this small adjustment to Hal's summary-and-reconstruction procedure, Hal once again goes through the same entire album, summarizing and reconstructing every photograph, this time using the adjusted summary-and-reconstruction procedure. Hal's optimizer then reviews the results and calculates another very small adjustment guaranteed to make Hal's next summary-and-reconstruction run on the same album produce slightly more accurate reconstructions, and the entire process repeats itself. The process typically loops for several million rounds, concluding when Hal arrives at a procedure that can't be improved by any very small adjustments. This process of iterated small adjustments, technically known as gradient descent, is of limited interest to us, since the logic of the gradient descent method itself does not say much about the nature of what an autoencoder ends up learning by employing gradient descent. One caveat worth mentioning, however, is that any concept, structure, or skill learnable through gradient descent must be 'soft'—that is, difficult to describe using explicit rules or formulas, but amenable to intuitions, heuristics, and approximations. (E.g. tennis is soft and arithmetic is not, cooking is soft and baking is not, 'go' is soft and chess is not, and linguists disagree on whether syntax is soft.) We will revisit the matter of the 'softness' of the structures that autoencoders learn a great deal over the course of this dissertation, basing our discussion not on the implications of the gradient descent method of learning but rather directly on the nature of the 'summarize-and-reconstruct' task, and ultimately even on a strong formal

analogy between the structures an autoencoder must learn and some canonically hyper-soft structures of literary-theoretic fame such as Martin Heidegger's 'moods' (Stimmung), Sianne Ngai's 'tones,' and Raymond Williams's 'structures of feeling.' Nevertheless, it may well be important to remember that the humble mechanistic origins of an autoencoder's knowledge shouldn't lead us to expect that what it learns would be itself rule-driven, simple, inflexible, reductive, or even rigorous in character, but just the opposite.

Let us recap, this time allowing ourselves to introduce some basic technical vocabulary when appropriate: Let 'Hal' be an autoencoder algorithm. Hal's input channel is a receptor for some predetermined type of digital media file. The type or types of file Hal can receive as input will depend on the design decision of the AI researcher who built Hal, but the most typical choices are receptors for digital images, receptors for digitized audio, and receptors for word processor documents. Whenever Hal receives an input media file x, Hal's output channel #1 outputs a short list of short numbers that we call *Hal's feature values for* x, and Hal's output channel #2 outputs a media file we call *Hal's projection of* x. We call the computation that determines output #1 Hal's *feature function*, and the computation that determines output #2 Hal's *projection function*. Throughout this dissertation, we will often think of Hal's *feature function* as Hal's 'worldview' or 'conceptual scheme,' and of Hal's *projection function* as Hal's 'imagination' or 'mimesis.' The technical relationship between Hal's feature function and projection function is as follows: Hal's projection function is a composition of Hal's feature function and of a 'decoder' function that, for every input x to Hal, receives the output of Hal's feature function (aka Hal's *feature values for* x) as its input and then outputs Hal's *projection of* x. In other words, *Hal's Projection Function (x) = Hal's Decoder Function (Hal's Feature Function (x))*. Finally, Hal has an optimizer mechanism, which can be turned on or turned off. We call the set of all the inputs Hal receives while its

optimizer is turned on Hal's *training set*, and call any input Hal receives while its optimizer is turned off *test data*. When Hal is *training*—that is, when Hal's optimizer mechanism is turned on—Hal's optimizer mechanism does the following: For every input x included in Hal's training set, Hal's optimizer mechanism compares x to Hal's projection of x and computes a quantity called Hal's *reconstruction error on* x or Hal's *loss on* x. Using a formula called gradient descent, Hal's optimizer mechanism then makes a small change to Hal's projection function that slightly reduces Hal's *mean error* on the training set—that is, slightly reduces the average size of Hal's reconstruction errors. When the optimizer alters Hal's projection function, it necessarily also (by logical entailment) alters the two functions that compose Hal's projection function: Hal's feature function and Hal's decoder function. For reasons that directly follow the artificial neural network architecture of autoencoders, and which our abstract model will replace with stipulation, *Hal's optimizer always alters Hal's feature function and decoder function in roughly symmetrical ways.* A trained autoencoder's decoder function is therefore roughly, and often exactly, just its feature function running in reverse: Hal's decoder function translates short lists of short numbers into media files by mirroring the steps Hal's feature function uses to translate media files into short lists of short numbers. Hal's projection function, therefore, is a matter of using Hal's feature function to translate a media file into a short list of short numbers, and then running Hal's feature function in reverse to get a media file again. Of course, since the variety of possible media files is much wider than the variety of possible short lists of short numbers, something must necessarily get lost in the translation from media file to feature values and back. Many media files translate into the same short list of short numbers, and yet each short list of short numbers can only translate back into one media file. This means, in an important sense, that Hal's projection function always replaces input media file x with 'stand-in' media file y: unless x happens to

be the exact media file that Hal's decoder function assigns to the feature values that Hal's feature function assigns to x, Hal's projection of x will not be x itself but some media file y that acts as the stand-in for all media files that share its feature values. The technical name for the set of all the media files that Hal uses as stand-in—that is, all the possible outputs of Hal's projection function—is *the image* of Hal's projection function. In this dissertation, we will often think about *the image of Hal's projection function* as Hal's *canon*—as the set of objects that Hal uses as the measure of all other objects.[6] Importantly, because of the symmetry between the optimization of the feature function and the optimization of the decoder, the logic by which Hal determines which of the media files that share the same feature values gets is 'stand-in' for the group is utterly inseparable from the logic by which Hal determines the assignments of feature values in the first place—or, in conceptual terms, the logic that determines Hal's canon is inseparable from the logic that determines Hal's worldview.

1.2 The Meaning(s) of Autoencoders

Does our dear Hal tell us anything remotely universal about canon, worldview, and mimesis, or are the structural relationships at play in Hal only the artifacts of a particular mechanical design? And even if Hal does provide us with some kind technoscientific ground for the longstanding literary-theoretic notion that canon, worldview, and mimesis are intrinsically bound to each other, can Hal tell us anything genuinely new about the nature of this bond?

[6]In the introduction to this chapter, I suggested that our formal analogue of canon would be a trained autoencoders lossless inputs. Formally speaking, the set of an autoencoder's lossless inputs is simply the image of the autoencoder's projection function. In later stage of this chapter, however, we will start finding great utility in sometimes explicitly thinking of the image of an autoencoder's projection function as the set of the autoencoder's 'lossless inputs'—the set of inputs that the autoencoder's projection function reconstructs without anything getting lost in the translation from input to feature values and from feature values to reconstruction.

In the preceding section, we devoted all of our attention to describing the mechanical procedures that Hal follows, saying nearly nothing of the abstract structures Hal's procedures are designed to channel. The present section will attempt to step beyond those limits, building up towards a theoretical interpretation of autoencoding called 'the manifold perspective.' The manifold perspective on autoencoding, closely identified with AI luminaries Bengio and LeCunn, proposes that what a trained autoencoder truly learns is a **space**, called the autoencoder's *manifold*, and all the facts about a given trained autoencoder follow from the form of this space. In terms of our literary-theoretic interest in autoencoders as potential exemplars of a structural relationship between worldview, canon, and mimesis, the potential importance of the manifold perspective is twofold: Firstly, the manifold perspective should allow us to stop thinking of the symmetry between an autoencoder's feature function and decoder function—the symmetry which binds worldview, mimesis and canon to a single learning process—as a contingent matter of mechanical design, and instead start thinking of the abstract structure this symmetry represents. This transition, if successful, will be our critical step towards establishing autoencoders as plausible windows into abstract matters of worldview, canon, and mimesis, rather than just windows into some contingent engineering choices. Secondly, the manifold perspective introduces a new formal concept—an autoencoder's *manifold*—to the company of such concepts as an autoencoder's feature function, an autoencoder's projection function, and the image of an autoencoder's projection function. If an autoencoder's feature function really is the equivalent of a worldview, an autoencoder's projection function really is the equivalent of (a method of) mimesis, and the image of an autoencoder's projection function really the equivalent of a canon, then what is the equivalent of this abstract space whose form supposedly implies worldview, mimesis, and canon all at once? One hopes, of course, that in discussing the math of autoencoders we're discovering

an entirely new way to articulate the logic of the worldview/mimesis/canon triplet, and in some respects we will in fact insist that 'manifold' is irreplaceable. Yet, on the other hand, this 'manifold' must be at least a cousin to the thousand and one concepts that contemporary affect studies—that 21st century hybrid of Phenomenological and Marxian literary criticism most closely identified with Lauren Berlant and Sianne Ngai[7]—keeps around exactly to discuss elusively holistic structures at the fountainhead of (among other things) worldview, mimesis, and canon: concepts like Althusser's 'ideology,' Bourdieu's 'habitus,' Heidegger's 'mood,' Williams's 'structure of feeling,' Ranciere's 'sensorium,' Simmel's 'forms of experience,' or Ngai's 'tone.' In Chapter 2 we will, naturally, grab both horns of the dilemma, proposing that the literary-theoretic concept of a manifold cannot be paraphrased away using familiar concepts, but nevertheless has a rich and tractable relationship to the conceptual repertoire of affect studies.

Because the manifold perspective on autoencoding promises to do so much for our purposes, we will not introduce the manifold perspective excathedra. Instead, we will try to impute the manifold perspective ourselves by looking for a universal element in the mechanically defined autoencoder algorithm we already introduced. To this effect, we will start by describing an autoencoder as per our previous definition of Hal, only this time Hal's algorithm won't be carried out by a machine but rather by a pair of humans playing an unusual collaborative game: Imagine that two expert art restorers specializing in Classical Greek pottery learn that an upcoming exhibition will unveil a previously unseen collection of Classical Greek pots, and they decide to put their understanding of Classical Greek pottery to an extreme test. The self-imposed rules of their game dictate that they must try to recreate the

[7]I would distinguish 'affect studies,' a tradition of literary and cultural studies marked by an interest in cultural-material systems as (in part) phenomenological systems, and 'affect theory,' a tradition of literary and cultural studies marked by an interest in the ontology of emotion.

pots from the new exhibition, but only Expert #1 may see the exhibition, and only Expert #2 may sculpt the recreations. Expert #2 will thus have to rely entirely on Expert #1's descriptions of the pots, and Expert #1 will have to rely entirely on Expert #2's ability to divine the original pots from her descriptions. Not yet content with the rules of their test, the two Classical pottery experts add one last constraint: instead of describing the pots in as much detail as she pleases, Expert #1's communications from the exhibition will be limited to written messages of 100 characters per pot. In the division of labor fixed by the rules of their game, then, Expert #1 will play the role of an autoencoder's **feature function**, and Expert #2 will play the role[8] of an autoencoder's **projection function**, with each pot in the exhibition acting as an input. (Quick reminder: 'feature function' is the formal term for what we have informally called an autoencoder's method for summarizing inputs, and 'projection function' is the formal term for what we have informally called an autoencoder's method of reconstructing inputs.) While this might seem a strange or pointless game for experts in Classical Greek pottery to play, I would propose that it has special merit as a test of our experts' grasp of Classical Greek pottery as a full cultural-aesthetic system: Most crucially, because 100 written characters cannot suffice for a naïve detailed description of a Classical Greek pot, our experts must invent a shorthand that relies on their grasp of the **grammar** of Classical Greek pottery—their (largely implicit) grasp of the constraints and logic of the variation between one Classical Greek pot and another, from the correlations and dependencies between various ceramic techniques, thematic motifs, ornamental patterns, and laws of composition, to the interactions between these tangible variables and

[8]More formally, Expert #2 plays the role of an autoencoder's decoder function, and it is the chain of Expert #1 reporting and Expert #2 reconstruction that plays the role of an autoencoder's projection function. The less precise formulation is helpful, though since we will mostly be thinking about inputs in their original state, inputs after they have gone through Expert #1's reporting—that is, have gone through the feature function—and inputs after they have gone through Expert #1's reporting and the Expert #2's reconstructing—that is, have gone through the projection function.

a pot's painterly and sculptural gestalt.

For the remainder of this section, we will rather scrupulously delve into the meaning and mechanics of this hypothetical test of our experts' gestalt understanding of the cultural-aesthetic system of Classical Greek pottery, with the intention of beginning to conceptualize the epistemic objects of autoencoders—in other words, establishing just what it is you know when you know how to summarize-and-reconstruct. Relatedly, we will be bracketing the issue of autoencoders' optimizer mechanisms (aka their training process) for the moment, maintaining our focus in this section on just ***what*** autoencoders learn rather than how they learn it. We will start off, then, by positioning our Classical Greek pottery experts to be the analogue of an already trained autoencoder—an autoencoder whose optimizer has been turned off after an epoch of training, locking in place the present version of the algorithm's summarize-and-reconstruct procedure. We will imagine, therefore, that our Classical Greek pottery experts have completed all their preparations for the test, having devised, coordinated, and extensively practiced their Classical Greek Pottery shorthand, and they are now committed to employing the resulting method in the test. Importantly, we mustn't assume that because the state of our experts after they devised their method of summary-and-reconstruction is analogous to a trained autoencoder, the state of our experts before they devised their summary-and-reconstruction method is any way analogous to that of an **untrained** autoencoder. For one thing—and sufficiently—among many other disanalogies, the work that our experts must perform in order to devise their method of summary-and-reconstruction is minimal compared to the learning process of an autoencoder, since our experts already possess a gestalt understanding of Classical Greek pottery.

The above limit on the scope of our analogy is, in my view, all for the better. Instead of speaking about experts who already gained a gestalt grasp of Classical Greek pottery

by whatever means, we could have described students of Classical Greek pottery who train their gestalt grasp of Classical Greek pottery as an autoencoder would, by repeatedly testing and revising a summary-and-reconstruction method for Classical Greek pots, testing their method, and revising it, but tying our experts' knowledge to autoencoding from the get-go in this manner would in fact defeat the purpose. Our goal is to establish that autoencoding demonstrates structural forces whose applicability goes far beyond explaining algorithms that explicitly use the autoencoder learning procedure, and which bear on the general phenomenon of grasping a domain's systemic grammar. We therefore want to start with agents that are not by definition anything like an autoencoder, and who paradigmatically exemplify having a gestalt grasp of a domain's systemic grammar, and argue that one rigorous way to discuss the structure of this 'gestalt grasp' is to construct the formalism of an epistemically equivalent trained autoencoder. In other words, we want to temporarily dissociate the concept of trained autoencoders from actual autoencoder machines, and instead pitch it as a general schema for the mathematical representation of an agent's gestalt grasp of a domain's systemic grammar. Presently, this means defining our agents simply as experts that posses a gestalt grasp of a domain's systemic grammar, and then arguing that a summary-and-reconstruction task that requires our experts to act as their epistemically equivalent trained autoencoder in fact calls on the entirety of our experts' gestalt grasp of their domain as no other activity would. If this proposed relationship between summary-and-reconstruction and gestalt understanding holds as promised, then thinking about the structural forces at play in summary-and-reconstruction might be a source of insight into the elusive but frequently crucial idea of gestalt understanding of the systemic grammar of a domain of objects—an idea that comes into play whenever we as literary and cultural scholars want to speak about, say, the cultural logic of late capitalist art, the style of German Baroque music, or the Geist

of the building and streets of 19th century Paris. Furthermore, if we can make good sense of the idea that the activity of summary-and-reconstruction is a uniquely thorough demonstration of one's gestalt understanding of the systemic grammar of a domain, we cannot be far off from arguing that some form of 'summary-and-reconstruction' is at play when humans produce facsimiles (reconstruction) of the world that seem to powerfully embody their subjective take on the systemic grammar of the world (summary). Guilty as charged, we will devote Chapter 2 of this dissertation to discussing works of literature as something like 'reverse autoencoders.'

Putting these promises aside for now, let us return to our Classical Greek pottery experts and their autoencoder vivant. Given that our Classical Greek pottery experts are required to perform the same work as a trained autoencoder, what summary-and-reconstruction method could our experts use to do so? While there may be multiple ways to operationalize a gestalt grasp of Classical Greek pottery, one strategy we know to be effective is for our experts to devise a *feature function*: Let our Classical Greek pottery experts devise a list of 100 descriptive statements whose truth varies strongly from one Classical Greek pot to another (e.g. 'this pot depicts war,' 'this pot's ceramic technique is uncommon,' 'the pot's composition is symmetrical,' 'this pot's affect is mournful,' 'this pot depicts worship'), and agree that on the day of the exhibit Expert #1 will fill out her message about each pot she examines with 100 numerical grades ranging from 0 to 9, marking 'strongly disagree' to 'strongly agree' for each statement. If our experts can devise a list that optimally complements their gestalt grasp of Classical Greek pottery then the resulting summary-and-reconstruction process will be logically equivalent to a trained autoencoder of the same abilities. What we can get from our Classical Greek experts hypothetical, then, is a top-down perspective on a trained autoencoder: rather than looking at our trained autoencoder as the outcome of a certain

formal learning process applied to a domain, we're looking at our trained autoencoder as a formal expression of an agent's gestalt grasp of a domain. By understanding the relationship that our experts' list of 100 descriptive statements—the feature function of their trained autoencoder—has to bear to their gestalt of Classical Greek pottery to serve as an effective summary-and-reconstruction method, we have an opportunity to understand something of what a trained autoencoder's feature function 'means' in general.

What should we make of the idea that our experts' 'feature function'—that is, our experts' list of 100 descriptive to grade from 0 to 9—is a formal expression of their 'gestalt grasp' of domain? At first blush, the idea of a list of 100 statements that implements a gestalt grasp of the systemic grammar of Classical Greek pottery can, and perhaps should, strike us as suspicious. A 'gestalt grasp of the systemic grammar of Classical Greek pottery,' after all, is just the sort of soft, ambient thing no list should ever manage to spell out. Crucially, it turns out that a list that implements our experts' grasp of the grammar of Classical Greek pottery in the relevant does not need to *spell out* a grammar of Classical Greek pottery, but rather to implicitly exploit the inferential powers latent in our experts' grasp of the grammar of Classical Greek pottery. One useful way to think about our experts' list-making practice, in this context, is to call on our familiarity with a real-life game of compressed communication and holistic knowledge: the social game '20 Questions.' Let's call a game of '20 Questions' scaled up to 100 questions a game of '100 Questions.' Speaking informally, we might think of our experts' list as an optimal set of questions for a game of '100 Questions' about Classical Greek pots ***where questions have to be submitted in advance***. In fact, the reader of this chapter should be able to acquire some first-hand experience with the idea of a list of questions expressing a systemic grammar by devising her own strategies for an imaginary game of '100 Questions Submitted in Advance' on some personal favorite domain such as

'20^{th} century novels' or 'R&B songs.' What makes a list of questions for a game of '100 Questions Submitted in Advance' effective? In an ordinary game of '100 Questions', we are always looking for the most relevant question we can ask in light of all the answers we received so far. In '100 Questions Submitted in Advance,' we are looking for a list of questions that each remain relevant no matter the answers to the other questions. To put it simply, we don't want our individually good questions to step on each others' toes and end up giving us redundant answers. Speaking more formally, this means that we need our list to consist of questions whose answers are both individually unpredictable and statistically independent from each other—and that, returning to our Classical Greek pottery experts, the list of 100 descriptive statements Expert #1 will grade from 0 ('strongly disagree') to 9 ('strongly agree') to record her impressions of a pot must consist of statements whose validity varies widely and independently from Classical Greek pot to Classical Greek pot. The questions on the list are, in this sense, the 'fundamental' questions about a Classical Greek pot from our experts' gestalt point of view: if Expert #1 examines some Classical Greek pot x and sends Expert #2 a message answering 90 of the 100 questions, then Expert #2's gestalt grasp of the Classical Greek pottery system can't help her estimate the values of the 10 leftover answers about x, and different random guesses at the values of the missing answers will lead Expert #2 to create pots that are wildly physically and visually different to each other from a 'naïve' point of view, even if they are identical in 90 out of 100 fundamental abstract respects from the expert's point of view. While this is well and good, and it is not hard to imagine that a list of questions that are 'fundamental' from some point of view therefore 'express' it in some interesting sense, the fact that our 100 questions must give unpredictable and mutually independent answers (i.e. unpredictable and independent grades from 0 to 9) when applied to Classical Greek pots can still seem rather indifferent as a supposed impetus

for thinking of the list as a formal expression of the viewpoint of a 'gestalt grasp of the systemic grammar of Classical Greek pottery.' In fact, this dialectical relationship will now have to get worse in order to get better: we will evoke a sense in which using the experts' list of questions to examine Classical Greek pots gives answers that 'obscure' the specificity of the domain of Classical Greek pottery, then argue that it's in this very sense that our experts' list expresses the *viewpoint* of a gestalt grasp of the systemic grammar of Classical Greek pottery.

Suppose that we choose to apply our Classical Greek pottery experts' optimal list of 100 questions to every Classical Greek pot in the world, recording each pot as a 100 numbers-long list of grades from 0 to 9. If our list indeed consists of statements whose validity varies widely and independently between different Classical Greek pots, it would mean that the set of all the accounts of individual Classical Greek pots we have produced using our grading system—a set which, since it contains an account of every Classical Greek pot in the world, is in some sense our account of Classical Greek pottery in total—is itself bereft of any structure, pattern, grammar or gestalt. The math, for what it's worth, checks out, but does it make conceptual sense that the list of questions that expresses our grasp of a domain's systemic grammar has to also be the list of questions that 'dissolves' the domain's systemic grammar? Here, as elsewhere, some habits of thought familiar to us from literary theory might help us understand what's going on: Recall the common critical-theory warning that systems of thought always 'naturalize' the worlds to which they're best attuned, such a great deal of literary, philosophical, and political thought misrecognizes worlds that are highly particular and structured—the world of male experience, the world of white middle-class families—as universal, average, neutral, unconstrained or typical. While one might be reluctant to call our experts' list of 100 descriptive statements to grade 0 to 9 a 'system of thought,' it

certainly naturalizes the domain of Classical Greek pottery. Indeed, I would suggest that our experts' descriptive method literally ***internalizes*** the domain of Classical Greek pottery to the extent that it's successful. Recall the set we have formally called *the image of an autoencoder's projection function*, and informally called an autoencoder's canon: the set of all potential outputs of the autoencoder's projection function, or equivalently, the set comprising the autoencoder's 'representative' object for each potential output of the feature function. In our Classical Greek pottery case, where we treat our two experts and their list of 100 statements as a trained autoencoder, the image of the autoencoder's projection function is the set we get by taking every possible list of 100 numbers between 0 to 9 and then replacing each numerical list with the object Expert #2 would make if Expert #1 were to send her that numerical list in a message. Seen in relation to the universe of all the objects Expert #2 could construct at the behest of messages, each numerical message from Expert #1 is like a treasure map instructing Expert #2 how many steps to take in each of this universe's 100 directions in order to arrive at the right spot. Indeed, for Expert #2 the meaning of a given message of 100 numbers between 0 to 9 is of the form: go zero steps in the direction of depicting worship, then five steps in the direction of symmetrical patterns, then nine steps in the direction of unusual ceramic technique, then three steps in the direction of depicting warfare... On this line of interpretation, then, our experts' list of 100 descriptive statements is a coordinate system for the space of possible Classical Greek pots, or at the very least a coordinate system for our experts' mental model of the space of possible Classical Greek pots. (Indeed, in mathematical parlance our space of Classical Greek pots is a 100-dimensional space, and each 'question' is a dimension of this space.) If this interpretation holds—which, granted, we have yet to fully demonstrate—it follows, with a little effort, that our experts' optimal descriptive method ***internalizes*** the domain

of Classical Greek pottery, in at least the following sense: to the extent that our experts' gestalt grasp of systemic grammar of Classical Greek pots is accurate, our experts' optimal descriptive method describes objects only in terms of their relative position in (or to) a space of all possible Classical Greek pots arranged in compliance with Classical Greek pottery's systemic grammar.

1.3 Internalizing Spaces

In his excellent exposition to Heidegger's concept of mood ('Stimmung'), Jonathan Flatly writes that 'any orientation toward anything specific requires a presumed view of the total picture, a presumption that is usually invisible to us—that is just the way the world is.' Recall that in our Classical Greek pottery experts' descriptive system, Expert #1—the 'feature function' of the trained autoencoder—translates whatever object she's describing into coordinates for a point in Expert #2's space of possible Classical Greek pots. Thus in relation to Expert #1's descriptive system, Expert #2's[9] space of possible Classical Greek pots is literally 'the 'precondition' and 'medium' (Heidegger) that 'makes it possible in the first place to orient oneself toward individual things' (Heidegger): when Expert #1 describes an object, her description simply is the coordinates of a location in Expert #2's space of Classical Greek pots. And, indeed, Expert #2's space of Classical Greek pots is very much 'a presumed view of the total picture' (Flatly), in the specifically Heideggerian sense in which 'total' must refer not merely to everything that is, but to everything that is possible and the relation of all possibilities to one another.[10] Expert #2's space of Classical Greek

[9]Recall that we are using Expert #1 and Expert #2 as analogues for the feature function and decoder function of an autoencoder, so neither expert 'determines' anything for the other in any hierarchical way. Rather, Expert #1's method of message-writing and Expert #2's space of Classical Greek pots are both the product of their cooperative planning and practice.

[10]See Ratcliffe 2013.

pots is a model—a 'presumed view'—of the totality of Classical Greek pottery: a model of what Classical Greek pots are possible and how the possibilities of Classical Greek pots relate to one another. And, indeed, the 'presumption' is, as Flatly says, 'invisible': Suppose that Expert #2's model of the totality of Classical Greek pottery is an imperfect model, simplified or biased compared to the real systemic grammar of Classical Greek pottery, and therefore that the reconstructions of Classical Greek pots that our experts' system produces simplify or distort the original Classical Greek pot. Clearly, we could never employ Expert #1's descriptive system to describe the difference between an original and reconstruction, since Expert #1's descriptive system assigns the original and reconstruction the same feature values. Much as we would expect, the descriptive system that presumes Expert #2's model of Classical Greek pottery can't register the structural impact of this model on our experts' reconstructions.

It should be fair to say, then, that in intuitive terms Expert #1's descriptive system is a worldview that apprehends individual things only in relation to the 'total picture' given by Expert #2's space of potential Classical Greek pots. What we are now hoping to show is that by giving an exact meaning to the idea that the 'total picture' that grounds a trained autoencoder is a space, we can also express exactly what it means to apprehend an individual thing in relation to a total picture. To begin, let us revisit the interpretation of our experts' list of 100 descriptive statements as a *coordinate system* for (a model of) the space of possible Classical Greek pots: Consider the fact that for any two pots x, y in Expert #2's universe of potential Classical Greek pots there is some series of edits that transforms a message specifying pot x into a message specifying pot y. Because the contents of a message are 100 numerical grades, this series of edits can be expressed as a series of addition and subtraction operations on the 100 grades specifying pots x. This means, in turn, that the series of

edits that transforms a message specifying pot x into a message specifying pot y can be represented by a 100 numbers-long list of its own, this time with numbers ranging -9 to 9: the list that we get by subtracting the grades specifying x from the corresponding grades specifying y. In formal terms, we're treating the messages specifying x and y as vectors, and then we subtract the vectors to receive their 'vector difference.' (While our discussion in this dissertation generally won't assume any familiarity with vectors, we will nevertheless take to saying 'x and y's vector difference' instead of 'the list that we get by subtracting the grades specifying x from the corresponding grades specifying y' from now on, for obvious reasons.) Now, if our experts really did devise the best possible list of 100 descriptive statements to use for their Classical Greek pottery summary-and-reconstruction—the list that best complements their gestalt grasp of the systemic grammar of Classical Greek pottery—then the vector difference between potential pots x and y provides a basis for expressing the holistic relationship between pots x and y in Expert #2's gestalt of Classical Greek pottery as a spatial relationship. Firstly, the vector difference between x and y lets us derive a quantity that, from a purely mathematical perspective, qualifies as the distance between two points respectively given by x's and y's messages.[11] If our experts' list of statements to grade is indeed effective, the distance between the two points respectively given by x's and y's messages will necessarily have a strong correlation with Expert #2's intuitive judgment of the overall (dis)similarity of x and y. Secondly, the vector difference between x and y lets us derive a rule for editing a message specifying pot x that mathematically qualifies as moving on a straight line drawn between the point given by x's message and the point given by y's message. In other words, the vector difference indicates the direction of the point given by y's message relative the point given by x's message. If our experts' list of statements to grade

[11] If a numerical list z qualifies as a vector difference between two points, the distance of the two points is given by raising every number in z to the second power, summing, and taking the square root.

is indeed effective, then moving along the straight line between x's message and y's message corresponds to gradually evolving pot x into pot y, each movement giving us a pot that is proportionally further evolved towards y-ness. Generally speaking, it is no great matter that we can coherently define Expert #1's messages as coordinates for points in space comprising all possible outputs. The vector difference between two points, however, isn't a quantity that we *define* to track the similarity between two objects or the transformation path between two objects: our only act of definition was the choice to treat the list of numbers composing each potential pot's message as Euclidean coordinates. The fact that elementary geometrical relationships between points in the space resulting from this act of definition seem to track conceptual relationships between pots in our experts' model of Classical Greek pottery, then, is an example of why one might want to say the messages 'really are' Euclidean coordinates.

While the above should go some way towards explaining why it makes good sense to treat our experts' list as a coordinate system for (our experts' mental model of) the space of Classical Greek pots, so far it can only offer us a very partial formal interpretation of the idea that this space determines our experts' descriptions of individual inputs. We know that our experts' descriptive system turns every input into coordinates for a point in the space of Classical Greek pots, and that it therefore describes Classical Greek pots in terms of their relative position in the space of Classical Greek pots, but what of inputs **outside** the domain of Classical Greek pots? Recall that if Expert #1 were to secretly apply her descriptive method to some Roman pot or to some Classical Greek statue, or even to secretly apply it to some Coca-Cola bottle, house, or car, when Expert #2 receives the message of 100 grades she would have no grounds to suspect anything is amiss, and she will simply produce a Classical Greek[12] pot that matches the parameters reported in the message. Indeed, this

[12]Pedantry would recommend referring to what Expert #2 produces as 'facsimile Classical Greek' pots, rather than Classical Greek pots. Readability would recommend to forego this pedantry.

facsimile Classical Greek pot will be 'identical' to the Roman pot, Classical Greek statue, Coca-Cola bottle or car in question as far as the properties that our experts' system measure go, though this 'identity' will now have very different implications depending on the case. How does the spatial interpretation of our experts' system apply to the transformation of a Roman pot, a Coca-Cola bottle, a potted plant, a Classical Greek statue, a house, or a car into coordinates for a point in the space of Classical Greek pots? Perhaps surprisingly, it applies just as naïve spatial metaphor might lead us to suppose. Intuitively, we expect that as a worldview that internalizes the domain of Classical Greek pots, our experts' descriptive method should give meaningful results when applied to objects that are 'close enough' to being Classical Greek pots, and give increasingly nonsensical results as we go further outside of its comfort zone: a Roman pot or Classical Greek statue should have quite a lot in common with their assigned Classical Greek pot, while a potted plant or Coca-Cola bottle will have a much cruder relationship to their assigned Classical Greek pot, and a house or car no meaningful relationship at all. Drawing on the same pool of intuitions, we might even imagine that what our experts' system does given an input that is not a Classical Greek pot is choose the Classical Greek pot that's 'closest' to the input. In order for these intuitions to apply, the space that constitutes our trained autoencoder's total picture—the space comprising all the objects in the autoencoder's canon arranged in accordance with the coordinates assigned by the autoencoder's feature function—has to be, from an alternative perspective, a section of a larger space. According to the theoretical approach known as the manifold perspective on autoencoders, the space that an autoencoder learns is just this kind of mathematical object: a manifold in the space of *all possible inputs*. The space of all possible inputs, formally called *input-space*, is the set of all possible inputs to an autoencoder, spatially arranged in accordance with their superficial properties as inputs rather than an

underlying systemic grammar. The exact details of an autoencoder's input-space depend on the exact way we define that autoencoder's input channel, but what matters for our purposes is that conceptually speaking input-space is the space we get by taking all possible inputs and arranging them in accordance with the most naïve possible axes of similarity and difference. A good paradigm case of an input-space is the space of all possible digital images, commonly called 'pixel-space': in pixel-space, every pixel is a dimension, and images are arranged according to the color values of their pixels. (Most points in pixel-space, it's well-worth noticing, are pure 'noise' images for most intents and purposes, whereas 'meaningful' images will typically flock to various lower-dimensional manifolds within the space.) The main thing that makes pixel-space a good paradigmatic case of a 'naïve' similarity space is that it's easy to see that we can reliably expect that any pair of e.g. photographs that are *extremely* close in pixel-space will also be close on a mature measure of similarity for photographs, and that the converse holds as well—any two photographs that are *extremely* similar in terms of objects, lighting conditions, angle, and so forth will be close in pixel-space. Because autoencoders use input-space distance between reconstruction and original to measure their own accuracy during the optimization process, this kind of *differentiable* relationship between input-space distance ('naïve' similarity) and 'substantive' similarity as measured within the domain of interest is a prerequisite for an autoencoding process.

We now, finally, get to directly talk about the concept of a *trained autoencoder's manifold*. A manifold, in general, refers to any set of points that can be mapped with a coordinate system. (Some manifolds can only be accurately mapped using a series of coordinate systems, rather than one global coordinate system, but even these latter manifolds can often be approximately mapped by a single coordinate system as per an autoencoder.) In the context of autoencoders, one traditionally uses 'manifold' in a more narrow sense, to mean a lower-

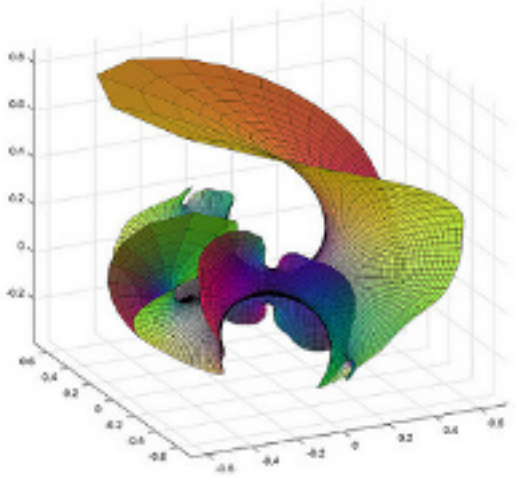

Figure 1.1: A 'two-dimensional' manifold within a three-dimensional space.

dimensional submanifold: a shape such that we can determine relative directions on said shape, quite apart from directions relative to the larger space that contains it. From the 'internal' point of view—the point of view relative to the manifold—the manifold in the illustration is a two-dimensional space, and every point on the manifold can be specified using two coordinates. From the external point of view—the point of view relative to the three-dimensional space—the manifold in the illustration is a ribbon-like shape curling hither and thither in a three-dimensional space, and every point on the manifold can only be specified using three coordinates.

Returning to our ongoing example of the Classical Greek pottery experts, we might reasonably ask to review the conjuration that transforms our experts' mental model of the field of Classical Greek pottery, which we recently described as a coordinate system for the space of Classical Greek pots, into a manifold in some kind of ceramic 'pixel space.' Recall that every point in our experts' coordinate system corresponds to one potential complete 100-digit message, composed of 100 grades between 0 and 9—one complete set of *feature-values*—

that Expert #1 might send to Expert #2. While there exist endlessly many hypothetical ceramic objects that Expert #1's 100-grades code cannot distinguish from each other, and in view of the inherent limits of all mental models it's extremely likely that their ranks include some pairs of real-life Classical Greek pots that Expert 1#'s code can't distinguish, we know that Expert #2's method of reconstruction must uniquely[13] assign every potential message of 100 grades between 0 and 9 a fully realized, concrete ceramic object. Speaking in our newly acquired spatial jargon, we might say that Expert #2's reconstruction method assigns every point in our experts' Classical Greek pottery coordinate system to a point in the raw input-space, marking the location of one Classical Greek pot in input-space. (That the set of all such input-space points should have a manifold shape follows from the assumption of a differentiable relationship between input-space distance and 'substantive' similarity.) A point in our experts' abstract coordinate system is, therefore, also a specific Classical Greek pot defined in input-space. Our Classical Greek pottery experts' descriptive method is the internal coordinate system of our experts' space of Classical Greek pots, describing each point in the manifold relative only to the manifold. When Expert #2 constructs a Classical Greek pot based on the specifications given in a message from Expert #1, however, she allows us to interpret the point specified by Expert #1's message from the external point of view, as a point in the space of all possible inputs. Indeed, by taking the set of all the reconstructions our experts can produce—what we have called the image of the trained autoencoder's projection function, or its 'canon'—and marking the location of each reconstruction in the space of all possible inputs, we get our experts' space of Classical Greek pottery from the *external point of view*, where our experts' space is a (multi-dimensional)

[13]Certain unsupervised deep learning algorithms outside of autoencoders, such as RBMs (Restricted Boltzmann Machines), rely on a probabilistic relationship between feature-values and output, for largely mechanical reasons. One may easily bridge the gap to the present discussion by speaking of the maximum likelihood output of a given set of feature-values where we speak of the output of a set of feature-values.

ribbon-like shape in the space of all possible inputs.

It's probably worthwhile, at this point, to insist on clarifying some potentially confusing issues about the relationship between the internal viewpoint of a manifold like 'Classical Greek pottery space' and the manifold as we see it from higher-dimensional, external points of view. Most important, the 100-dimensional space we're calling 'Classical Greek pottery space' isn't a space that we get by first finding a parameterization for the full input-space ('naïve all-logically-possible-ceramics-objects space') with, let's say, 1000 meaningful questions and then picking from among them the 100 questions that are most pertinent to Classical Greek pots. Instead, the 100 questions that act as the coordinates for Classical Greek pottery space may well only have meaning from the point of view of Classical Greek pottery space: the Classical Greek pottery expert doesn't necessarily employ a general concept of symmetry when judging 'is this pot symmetrical,' or use a general concept of worship when judging 'does this pot depict worship,' but rather makes Classical-Greek-pottery-specific versions of these judgments, which may well fail to track the relevant concept when used outside the domain of Classical Greek pottery. While this insistence might initially sound like an artifact forced on us by our analogy between our human experts and autoencoders, the view that judgments, and particularly expert judgments, are essentially embodied/embedded in a context rather than the application of a fully abstract universal judgments is an old favorite in Phenomenology (Heidegger, Merleau-Ponty) and social theory (Bourdieu, Williams). An expert in Greek pottery might well not know how to judge symmetry in a Chinese pot—if she cannot identify certain Chinese decorative patterns as patterns, for example, then she won't be able to tell whether the same patterns appear on both sides of the pot—or how to judge whether a Chinese pot depicts worship.

With the above in mind, we might want to consider two kinds of possible 'outside' points

of view on the dimensions of a space like 'Classical Greek pottery' space. First, there is the point of view of input-space itself—in this case, a space where all logically possible ceramics objects are arranged according to some 'naïve' organizing principle akin to pixel-space. Like pixel-space, what we might call 'naïve all-logically-possible-ceramics-objects space' does not comprise only real, typical, or sensible ceramic objects, but rather contains every logically possible ceramic blob or splatter, organized along several thousand dimensions that each measure an extremely local property. When we look at our experts' Classical Greek pottery space from the input-space's point of view, we're looking at how movement in Classical Greek pottery space translates to movement in the 'naïve' space of all possible ceramic blobs and splatters. Second, we can also conceive of the point of view of a third expert, an 'all-cultures-pottery' expert, and of a corresponding manifold within 'all-logically-possible-ceramic-objects space' (input-space) that has more dimensions than Classical Greek pottery space but fewer dimensions than the input-space. From the 'all-cultures-pottery' expert's point of view, the Classical Greek pottery experts' space is a 'limited' perspective on Classical Greek pots, because it cannot express, for example, all of the respects in which a given Classical Greek pot is different and similar to a given Chinese pot.

It could be tempting to think, at this point, that because Classical Greek pottery space is a lower-dimensional slice of the already low-dimensional (relative to input-space) 'all-cultures-pottery space,' the latter space is simply Classical Greek pottery space plus a few more dimensions, and therefore we can specify Classical Greek pottery space relative to 'all-cultures-pottery space' by choosing a certain fixed set of coordinates in the 'new' dimensions that restricts us to the Classical Greek pottery slice of 'all-cultures-pottery space' and leaving the 'original' dimensions free. In reality, this will practically never be the case: although the space of Classical Greek pottery is a subset of the space of all cultural pottery, the questions

that are best for navigating the space of Classical Greek pottery are not a subset of the questions that are best for navigating the space of all cultural pottery. To see why, imagine that we play a game of '20 Question Submitted in Advance' where it is given that the object is a marine mammal. Later we play a game of '40 Questions Submitted in Advance' where it is given that the object is an animal. Although one category is a subset of the other, the questions that you submit in the second game will not be the same 20 questions you submitted in the first game plus 20 new questions, but rather a new set of questions. Each space is a *totality* (to rudely borrow the term from Western Marxism) unto itself, and every aspect of analysis—that is, every dimension—has structure only relative to this totality. Still, we can reliably expect some interesting relationship between movement relative to 'all-cultures-pottery space' and movement relative to Classical Greek pottery space, at least where the two manifolds overlap: it's probable enough that some of the dimension of 'all-cultures-pottery space' will roughly correspond to *some* of the dimensions of Classical Greek pottery space around points where the two manifolds overlap. It would be plausible enough, for instance, to imagine that the question 'is the pot symmetrical' is a key question to ask both when describing a given Classical Greek pot relative to Classical Greek pots and when describing a given cultural pot of whatever kind relative to all cultural pottery. In this case, even though the two questions don't have strictly the same meaning (because different expertise give the question different embodied/embedded meaning), we would expect that if we take the coordinates of a given Classical Greek pot x in both spaces, then whether we make a small increase in the grade of the 'pot symmetry' coordinate of our current location in 'all-cultures-pottery space' or make a small increase the 'pot symmetry' coordinate of our current location in Classical Greek pottery space, we would arrive at roughly the same slightly-more-symmetrical-than-x Classical Greek pot.

Over the course our discussion so far, we allowed ourselves to bracket a distinction that will now become of key importance: the distinction between our experts' mental model of the field of Classical Greek pottery, on the one hand, and the ***actual*** field of Classical Greek pottery, on the other hand. Our experts' mental model of the field of Classical Greek pottery, we've seen, has the internal structure of a low-dimensional coordinate system, which we informally called our experts' 'Classical Greek pottery space,' and the external structure of a low-dimensional submanifold in input-space comprising all the input-space points corresponding to ceramic objects that Expert #2 can construct. (The set of input-space points that compose this manifold, we know, is the same set we have informally called a trained autoencoder's *canon*, or formally called *the image of a trained autoencoder's projection function.*) How should we think about the ***actual*** field of Classical Greek pottery within this formal context, then? If we could meaningfully represent the actual field of Classical Greek pottery (as distinct from our ideal trained autoencoder's 'mental model' of the field of Classical Greek pottery) *noumenally*, then the mathematical trope of a trained autoencoder would in fact be far less philosophically interesting than we propose, since in that case we'd have a separate mathematical trope for the field 'in itself,' suggesting that the transcendentally respectable subject's representation of the field should share a form with *that* mathematical trope rather than with the mathematical trope of trained autoencoder. An ideal trained autoencoder's model is, in this sense, as much noumenal representation as we have on offer. We *can*, however, meaningfully think of the actual field of Classical Greek pottery *phenomenally*, as simply the totality of true Classical Greek pots.[14] More cleverly, though very similarly for our purposes, we might think of the actual

[14]We might think of the set 'actual Classical Greek pots' empirically, as the set of all existing Classical Greek pots, or modally, as the set of all possible objects that would qualify as Classical Greek pots in some ideally extrapolated sense.

field of Classical Greek pottery as an (effectively) inexpressibly complex probabilistic 'data distribution'—a kind of weighted lottery that says, for every logically possible ceramic object, how common it is for the field of Classical Greek pottery to generate this object—that we can represent phenomenally by randomly sampling real-world Classical Greek pots to build a representative 'actual Classical Greek pottery' set. Our practical representation of the actual field of Classical Greek pottery as distinct from an ideal autoencoder's 'mental model' is, in this sense, constructed just like what we typically call the *test data* for a trained autoencoder: it is a set of inputs sampled from the same source as the *training set*, but withheld during training.

As we might logically expect once we consider that a trained autoencoder's manifold is the set of the trained autoencoder's input reconstructions, and that an autoencoder's training process is designed to minimize its reconstruction error—that is, minimize input-space difference between reconstruction and original—on any set of inputs sampled from the same source as the training set but withheld during training, there is a binding geometrical relationship between a trained autoencoder's manifold and (a fair sample from) the data distribution of its training domain. If our experts' space of Classical Greek pots is a set of input-space points composing a low dimension manifold in input-space, the *actual* field of Classical Greek pots is a set of input-space points composing a pointillist cloud around the contours of the manifold:

In an obvious though informal geometric sense, a trained autoencoder's low-dimensional input-space manifold is the *intelligible structure* of the high-dimensional (but condensed) cloud of data-points that makes up the autoencoder's real-world domain. The data-points of a real-world domain, we know, are distributed in a pattern all too complex to align with any trained autoencoder's manifold, but reference to an optimal trained autoencoder's manifold

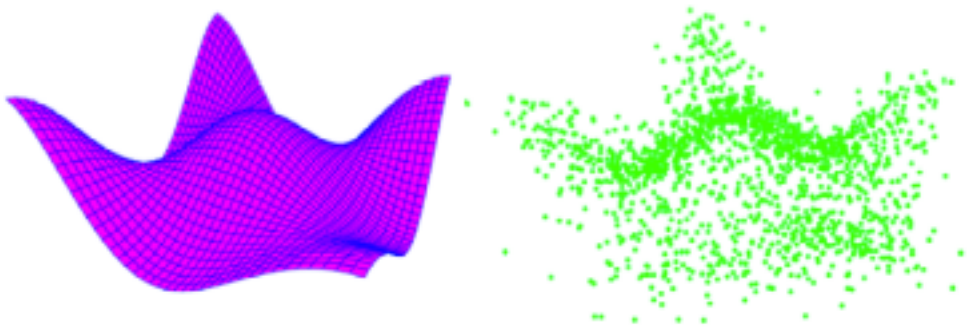

Figure 1.2

'explains' this pattern by breaking it down into an intelligible macroscopic aspect and an unintelligible microscopic aspect: an intelligible low-dimensional manifold structure, and an unintelligible high-dimensional scatter pattern in a small radius around the manifold.

1.4 Reading a Manifold

Over the course of the previous section, we had several occasions to observe that the external structure of our experts' space of Classical Greek pottery—that is, the input-space manifold of our trained autoencoder—is directly specified by what we've called a trained autoencoder's canon, or more formally *the image of a trained autoencoder's projection function*: the set comprising all the reconstruction outputs that a given trained autoencoder can produce. In this final section, we will argue that in fact the set of inputs that compose a trained autoencoder's *canon* describes much more than just the 'shape' of our autoencoder's space from an external point of view. First, a trained autoencoder's canon also specifies a metric of *internal distance*, of distance between objects from the point of view of the autoencoder's own internal space, through an implicit quantity that we might call the input-space manifold's *density*. Second, a trained autoencoder's canon implicitly specifies a strong approximation

of the trained autoencoder's *projection function*, for all practically relevant inputs. Putting these two together, we also derive a strong approximation of the trained autoencoder's *feature function*, giving us a strong approximation of a given trained autoencoder's total functionality from its *canon* alone.

How does a trained autoencoder's canon give a metric of 'internal distance'? The reasoning at play here is a tad technically involved, but can be summarized as follows. Internal distance on a manifold is, in principle, defined independently of the manifold's appearance as a shape in input-space. In Figure 1.1, for example, the crisscrossing lines on the manifold depict a coordinate system that 'interprets' input-space distances between points on the manifold very differently in different regions of the manifold. (Every line-segment between intersection points in Figure 1.1 marks a distance of length 1 in the coordinate system assigned to the manifold, regardless of the input-space length of the line-segment). In other words, a given coordinate system can have a dynamically varying degree of sensitivity to input-space distances on the manifold: sometimes a small input-space distance between points on the manifold might translate to a large difference between their coordinates, and sometimes a large input-space distance between points on the manifold might translate to a small difference between their coordinates. Because an autoencoder's feature function is ultimately discrete—that is, there is a finite number of setting for each feature[15]—the dynamically varying sensitivity of a trained autoencoder's coordinate system to movements on the manifold is ultimately expressed as a dynamically varying threshold for registering a movement on the manifold as a change in feature values. This property of the autoencoder's feature function is, in turn, expressed by what we might call the varying 'density' of the autoencoder's manifold—visually, the varying density of the ribbon-shaped cloud of input-

[15]Both human brains and artificial neural networks are 'pseudo-continuous,' but can't keep making smaller and smaller distinctions literally ad infinitum.

space points formed by the autoencoder's canon. While the input-space distance between neighboring points in this ribbon-shaped cloud of input-space points will vary, we know that from the point of view of the autoencoder's coordinate system the distance between each pair of neighboring points in this ribbon-shaped cloud is always 'a single step.' This information about the autoencoder's coordinate system's dynamically varying sensitivity to input-space distances on the manifold provides us with what a mathematician would call the autoencoder's Riemannian metric on the manifold—its system for deciding the 'internal distance' between points—and effectively[16] determines the autoencoder's coordinate system. For ease of use, we will henceforth refer to the pointillist ribbon of input-space formed by a trained autoencoder's canon, which gives us both the shape and 'density' of the autoencoder's manifold, as the input-space *form* of the autoencoder's manifold.

As we can see (or take on faith), the 'internal' point of view on points within the manifold associated with an autoencoder is mathematically determined by the input-space form of the autoencoder's manifold. Given the canon that composes our Classical Greek pottery experts' space of potential Classical Greek pots, for example, we can mathematically deduce our experts' coordinate system for their space of potential Classical Greek pots.[17] The real benefit of understanding an autoencoder as a manifold, however, lies in the fact the input-space form of an autoencoder's manifold also allows us to make spatial sense of the trained autoencoder's treatment of inputs **outside** of its canon—that is, make spatial sense of the autoencoder's treatment of those inputs its projection function actually transforms in some way. The term 'projection' is, in fact, a shorthand for 'orthogonal projection to the

[16]Coordinate systems with the same Riemannian metric ascribe the same geometrical relationships between points.

[17]As we discussed in the beginning of this section, our experts' coordinate system for the space of Classical Greek pots in turn gives us our experts' model of the system-grammatically structured similarities and differences between individual potential Classical Greek pots, expressed as the vector difference between the internal coordinates of any two points x, y in the manifold.

manifold,' which means treating an input x as a point in input-space and taking the nearest input-space point covered by the manifold: if input x is itself on the manifold, the projection function outputs x itself, and if input x is not on the manifold the projection function output the 'canonical' point that is nearest x in input-space. In our Classical Greek experts hypothetical, we remained agnostic as to whether our experts' space of potential Classical Greek pots is a perfect match to the domain of Classical Greek pottery, and so we relegated the question of projection to the autoencoder's treatment of objects outside the domain of Classical Greek pottery—objects like Roman pots, Classical Greek statues, houses, cars. In reality, however, mental models are always simplifications or at least abstractions of the domain that they model, and therefore nearly all of a trained autoencoder's reconstructions of inputs from its own domain will involve a substantial projection. Indeed, it is the ubiquity of substantive projection that makes it worthwhile to conceptualize a trained autoencoder as a system of mimesis in the literary-theoretic sense: a system that represents objects using imitations whose relationship to the originals betrays a worldview.

While it is obvious that the input-space form of the autoencoder's manifold determines which point on the manifold is closest (in input-space coordinates) to any given point x, it may not be so clear what relationship this operation has to the 'internal' structure of the manifold, and by extension to the trained autoencoder's model of the system-grammatically structured similarities and differences between the objects that compose its canon.[18] A very helpful way to think about the meaning of the projection operation is as follows: Given an input-space point x that isn't covered by the manifold, let's take point y, the *furthest* input-space point from x that is covered by the manifold. Let us 'travel' on the manifold, starting

[18] As our reader might remember, it is given that distances in input-space are maximally naïve measures of similarity. Thus it may appear strange that a trained autoencoder should use distance in input-space as the rule of projection.

from point y, until we are as close as we can be (in input-space distance) to point x—that is, until we are at the projection point of x. From the internal point of view, every step we take along the path from y to the projection point of x corresponds to a system-grammatically meaningful change to y. Once we reach the projection point of x, however, no system-grammatically meaningful change to y will get us any closer (in input-space coordinates) to x. The projection point of x is thus the one point on the manifold whose input-space difference from x is **completely** inexpressible from the internal point of view. We cannot traverse any of the input-space distance between x's projection point and x by moving on the manifold—or, speaking from the internal point of view, we can't traverse any of the distance between x's projection point and x by making system-grammatically meaningful changes. Consequently, by determining an autoencoder's projection function, the input-space form of its manifold also determines its feature function's output on inputs outside the manifold: an autoencoder's feature values for x are simply the feature values (that is, the internal coordinates) of x's projection point. As an important corollary, because distance in input-space is an extremely naïve measure of similarity that is only sensible across short distances, a trained autoencoder's 'area of competence' is in fact the literal input-space area around its manifold. In other words, in order for an autoencoder's feature function to record system-grammatically meaningful properties of the input, and for its projection function to simplify or interpret (rather than simply distort) the input, the input has the be close enough to the manifold to render input-space distance meaningful.

Looking back at the total course of our discussion in this chapter, we should now be ready to sum up the major literary-theoretically suggestive points of our literary-theory-facing introduction to autoencoders. A *trained autoencoder*, we asserted, is an algorithm that takes any concrete object as an input, and assigns the input object two representations: an ab-

stract representation in the form of a point in the autoencoder's feature space, where a point specifies a constellation of structural or gestalt properties, and a concrete representation in the form of a 'stand-in' object from a set of concrete objects that we've called the trained autoencoder's *canon*. We called the computation that transforms an object into an abstract representation the autoencoder's feature function, and the computation that transforms an object into a concrete representation chosen from among the objects in the trained autoencoder's canon the autoencoder's *projection function*. The projection function of a trained autoencoder, we have argued, is a kind of method of mimesis, and the feature function of a trained autoencoder is a kind of model of the generative structure of a world, as well as a conceptual scheme that interprets objects and phenomena in accordance with a gestalt worldview predicated on this model.

What makes the concept of a trained autoencoder interesting to a literary theorist, we suggested, is that a trained autoencoder's 'method of mimesis' and 'worldview' are both logically interdependent aspects of the trained autoencoder's overall holistic structure, much as a human subject or human culture's process of mimetic reproduction and representation is closely related to a subject or culture's schemas for interpreting the world. While we have not yet tried to definitively associate the overall holistic structure of a trained autoencoder with one philosophical or literary-theoretic master term, over the course of our discussion chapter we found various occasions to at least provisionally describe a trained autoencoder as a kind of gestalt, a kind of generative model, a kind of mood, or a kind of cognitive map. (In an upcoming stage of our discussion, we will call on literary-theoretic help from Ngai, Heidegger, and Williams to inquire further into the interpretation of this structure, and to discover whether we might speak of 'autoencoder structures' structuring canons, worldviews, and mimesis in the realm of individual and/or collective human cognitive and

cultural activity.)

In the final section of this chapter, we learned that the holistic structure that determines both the feature function and the projection function of a given trained autoencoder is in fact given by the trained autoencoder's canon—that is, by the set of all the objects that a given trained autoencoder can use as concrete representations—in and of itself. Specifically, we learned that both the feature function and the projection function of a given trained autoencoder are effectively identical to basic geometric properties of an input-space manifold corresponding to the trained autoencoder's canon: internal distance on the manifold, and least input-space distance from relevant input-space points to the manifold. We called these geometric properties the input-space form of a trained autoencoder, and determined that the input-space form of a given trained autoencoder is effectively fixed by the list of input-space points (input objects) that compose the trained autoencoder's canon. The input-space form of a trained autoencoder is, in this sense, an organic, emergent group-property of the input objects that collectively make up the trained autoencoder's canon.

One reason that identifying an autoencoder with its canon is conceptually significant, we might observe, is that because the canon of a trained autoencoder is simply a set of concrete input objects, we can now start to identify the idea of autoencoding with a type of *set of concrete objects*. Some sets of concrete objects, we now know—the 'special' sets that are each the canon of a trained autoencoder—correspond to trained autoencoders,[19] and therefore correspond to methods of mimesis bound with worldviews and, at least provisionally, correspond to gestalts, generative models, moods, and cognitive maps. Furthermore, this 'autoencoder structure' is what we might call an *aesthetic* property of the canon in the Kantian sense, in as much as it's determined solely by the intrinsic *forms* of the objects

[19]We put an upper bound on the number of parameters in an autoencoder relative to the size of the set, to make sure that only those sets who are the canons of non-trivial autoencoders qualify as 'canons.'

that make up the canon, independently of any external material relationship between these objects in the real world or superimposed conceptual relationships. In fact, autoencoders are in general 'aesthetic' learners—they model the causal, generative order of a material system underlying the training data by modeling the aesthetic structure of the training data. It is a fundamental assumption of autoencoding that the aesthetic structure of a dataset can be expected to effectively express something of the causal, generative order of the material system that produced the dataset. It should be no surprise, in this sense, that a trained autoencoder's own cognitive system is itself accessible as an aesthetic structure in the data generated by the trained autoencoder. The set of objects that makes up a given trained autoencoder's canon, then, aesthetically—that is, by form alone—represents its trained autoencoder, so aesthetically represents a *method of mimesis* and a *worldview*. Furthermore, as a robust formal representation of a trained autoencoder, this set of objects constitutes a structure that we might want to describe (at least noncommittally) as an *aesthetically embodied* gestalt, generative model, mood, or cognitive map.

Chapter 2

Ideas in Things

The mathematical trope of autoencoding, we have seen, has many features certain to delight and educate the literary theorist. The topic of this chapter, and in fact of this dissertation as a whole, is the fateful conjunction of several of these features in the figure of what we have called a trained autoencoder's canon (formally, a trained autoencoder's *input-space manifold* or *the image of a trained autoencoder's projection function*): the set of all possible outputs for a given trained autoencoder, and equivalently the set of all objects that a given trained autoencoder can encode and reconstruct with zero loss. The canon of a trained autoencoder, we will argue, is a kind of constructed world that aesthetically embodies something close to a gestalt, a generative model, a mood ('Stimmung'), or a cognitive map, directed at the abstract structure of the real world. We argue that by virtue of these properties—by virtue of the world-like materiality of an autoencoder's canon, the gestalt-like meaning of an autoencoder's canon's aesthetic structure, and an autoencoder's canon's role as an ineffable cognitive schema for the real world—the overall function and logic of a trained autoencoder's canon represents a powerful example of 'ideas in things,' in a sense that takes after Williams Carlos Williams ('no ideas but in things') as much as after Aristotle. In other words, we argue that the canon of a trained autoencoder is a powerful example of a kind of thought-

structure or meaning that is necessarily aesthetically embodied, necessarily inseparable from the aesthetic resonance of a constructed world.

In Chapter 1, we argued that the canon of a trained autoencoder is a set of objects or phenomena whose aesthetic unity represents the systemic structure of a real-world domain, and serves as a conceptual scheme for interpreting real-world objects and phenomena, effectively equipping ourselves with a kind of 'romantic theory of the Symbol'—or at least a theory of the romantic Symbol's Marxian niece, 'aesthetic of cognitive mapping'—for curated sets of objects or phenomena. Building on the above interpretation of the formalism of a trained autoencoder's canon, this chapter will propose that certain works of literature—possibly every work of literature, to some extent—are functionally analogous to a trained autoencoder's canon. This generalized concept of a trained autoencoder's canon, we will argue, is suggestive of an information-theoretically strong sense in which a work of literature might aesthetically embody something close to a gestalt, a generative model, a mood, or a cognitive map, directed at the abstract structure of the real world. Furthermore, to the extent that a work's meaning is analogous to an autoencoder's canon, the meaning of a work will start exactly where distinctions between form, structure, and content give out, putting us in contact with the metaphysics of strongly romantic, Symbolist, or Modernist ideas of literary meaning that impute to the artistic object a radical immanence. The technoscientific burden that our mathematical trope of a trained autoencoder's canon threatens to impose upon the constitution of a work of art, then, is in effect convergent with a doctrine of artistic immanence (the immanence of experience, of structure, and of sense to one another in a true encounter with a work of art) right out of 'Minima Moralia':

> 'Those who have subjected themselves in earnest, out of the energy of their precise reaction, to the discipline of a work of art, to the compulsion of its shape,

> of its immanent law of form, find the objection against what is merely subjective in their experience dissolving like a threadbare appearance [Schein], and every step they take further into the matter, by virtue of their extreme subjective innervation, has incomparably greater objective power than comprehensive and much-vindicated conceptual formations, such as that of 'style,' whose scientific claim comes at the cost of such experience.' (70)

This chapter makes the case that at least certain works of literature in fact aesthetically embody something close to a gestalt, a generative model, a mood, or a cognitive map in a strong information-theoretically formal sense related to autoencoding, and that a literary work's production or instantiation of this information-theoretic property powerfully elides a work's aesthetic texture, abstract logic, and mimetic content. What kind of works of literature, and in what capacity? We take this question up with due literary-historical, and literary-theory-historical, seriousness in Chapter 3, but one particularly useful, if particularly crude, reference point we might use to begin is the Symbolist paradigm of literary practice, which is nowadays a little faded as a taxonomical criterion but contemporary critics took as a common framework of *Les Poètes maudits*, Flaubert, Yeats, Valéry, Eliot, Proust, Joyce, and Stein, among others. The Symbolists, if there were Symbolists, believed that literature can esoterically represent, by means of an arrangement of concrete materials, primordial Ideals that underlie reality, which are the reason that reality is full of hidden correspondences. One lesson from the mathematical equivalence of a trained autoencoder and its canon, we propose, is that an appropriately selected set of concrete objects can, all by itself, aesthetically express a method of mimesis and a worldview, and thereby constitute something like a gestalt, a generative model, a mood, or a cognitive map.[20]

[20]The above may, or may not, have a meaningful relationship to the Deleuzian idea of an 'assemblage.' See Daniel for a good example of the concept's application in a related capacity.

In the Symbolist manifesto, arguably the first canonically Modernist *ars poetica*, Jean Moréas writes: 'In this art, scenes from nature, human activities, and all other real world phenomena will not be described for their own sake; here, they are perceptible surfaces created to represent their esoteric affinities with the primordial Ideas.' (Klein)[21] On the view that this chapter will propose, a literary work often comprises or constructs a set of objects or phenomena that, like a trained autoencoder's canon, constitutes something close to a gestalt, a generative model, a mood, or a cognitive map. All works of literature, we would propose, more or less saliently engage in the act of imaginatively demarcating a set of objects or phenomena, in at least the following obvious but important sense: the fictional world or world of imagery that a literary work traverses in its narrative, rhetorical, or lyric process remains in the wake of a work's temporal arc as a kind of archive, a curated set comprising all the objects or phenomena spanned by the literary work. This literary act of demarcating a set of objects or phenomena, we will argue—and in doing so suggest that something like a Modernist's 'ideas in things' is a concretely possible form of cognition—can plausibly rely on the same information-theoretic principles that underlie autoencoding to aesthetically constitute an 'idea' about the structure of the real world, and an expression of a subject's orientation toward reality. This information-theoretic aspect of this chapter's argument, it's worth observing, will rely not on a broad analogy but on deduction in a formal model, together with considerations as to the adequacy of the model.

On a more subtle level, the discussion in this chapter will also begin to set the groundwork for a reintegration of certain broadly romantic literary-critical (or, really, metaphysical) concerns that motivate the argument above—one doesn't, after all, just wind up mathemat-

[21] In the original: 'Ainsi, dans cet art, les tableaux de la nature, les actions des humains, tous les phénomènes concrets ne sauraient se manifester eux-mêmes ; ce sont là des apparences sensibles destinées à représenter leurs affinités ésotériques avec des Idées primordiales.'

ically reifying the romantic, Symbolist, Modernist theology of special thought and meaning inextricable from the aesthetic resonance of an imaginative artifact by accident—into the broadly 'cultural materialist' tradition of contemporary literary theory, and in particular the Marxian/Phenomenological tradition (Lukács, Jameson, Flatley, Berlant, Ngai) that gently informed us throughout Chapter 1. The proposition that some literary works are functionally analogous to the canon of a trained autoencoder, we will argue, can provide a powerful foundation for a new kind of inquiry into the logic of some of the more philosophically elusive literary-theoretic concepts of ideological-affective content, from a literary work's 'structure of feeling' (Williams) or 'cultural logic' (Jameson) or 'sensorium' (Rancière), in a manner that illuminates their application both in literary criticism and in cultural theory at large. In her groundbreaking monograph on affect-theoretic literary criticism, 'Ugly Feelings,' Sianne Ngai discusses the key role that a concept of 'tone' must play in the foundations of aesthetically sensitive 'cultural materialist' or 'ideological' literary criticism, observing that 'in its generality and abstractness [...] the concept of 'tone' seems ideally suited for the analysis of ideology,[22] which, as the materially embodied representation of an imaginary relationship to a holistic complex of real conditions, clearly shares tone's virtual, diffused, but also immanent character.' (47)

Whatever deeper bond it is that brings together the ideas of 'tone' and 'ideology' as above, we'd like to say, this bond may well extend to introduce the mathematical idea of a trained autoencoder's manifold to the company of this pair. A trained autoencoder's *manifold*, we argued at the end of Chapter 1, is the overall abstract structure of a trained autoencoder, specified in the implicit structure of the set of input-space points that we've called a trained autoencoder's 'canon.' Looking back on our intermittent ventures into Marxian and Phe-

[22]Ngai's concept of ideology, of course, implicitly takes after Althusser's canonical definition.

nomenological vocabularies to describe the meaning(s) of a trained autoencoder throughout Chapter 1 with Ngai's tone/ideology comparison in mind, we would readily find that Ngai's description of the common form of tone and ideology makes for a strikingly good summary of Chapter 1's interpretation of the mathematical trope of 'a trained autoencoder's manifold': a trained autoencoder's manifold per Chapter 1 is, much like Ngai's 'ideology,' a materially embodied representation of an abstract, general imaginary relationship to a holistic complex of real conditions, reminiscent of aesthetic structures like 'tone' in its virtual, diffused, but also immanent character.

While we will clarify, explore, and justify the above paradigm in greater detail throughout Chapter 3, it's worthwhile to point out certain particulars that bear directly on this chapter's proposition that a literary work can be functionally analogous to the canon of a trained autoencoder—that is, analogous to a set of input-space points that (effectively)[23] implicitly expresses the overall abstract structure of a given trained autoencoder. Early in Chapter 1, we offered to interpret the mathematical trope of a trained autoencoder as (a model of) the recapitulation of a material world's 'soft' systemic structure as the gestalt of a worldview, and started broaching the idea that the 'meaning' of a given trained autoencoder is a kind of mood or subjectivity latent in the material structure of a world. The canon of a trained autoencoder, we concluded at the end of Chapter 1, is an effective full specification of a trained autoencoder in the form of a constructed world, a privileged ensemble of 'made up' or specially curated worldly objects that effectively embody the autoencoder. Putting the two together, it's apparent that to theorize an information-theoretic property whereby a work of literature may be functionally analogous to the canon of a trained autoencoder is to

[23]When we describe a relation of logical entailment (e.g. identity, determination, or expression) between x and y as 'effective' in this dissertation, we mean to imply both that the relationship holds 'near enough'—that is, holds in all relevant respects, up to any desirable degree of precision—and that the relationship is computationally efficient in a strong informal sense, such that deriving y from x is computationally cheap.

theorize the work as a kind of aesthetic fossil-record of the structural recapitulations between subjectivity, imagination, and material reality: we will be looking at a literary work, at least in part, as an ensemble of imaginary objects whose aesthetic unity is the scheme of some totalizing worldview, which is in turn a worldview whose gestalt internalizes and naturalizes the systemic logic of some real material world. We theorize the literary work, in other words, as an aesthetic archive of some force or structure that is very closely tied to Ngai's version of 'ideology,' and possibly identical to Ngai's version of 'ideology' in certain cases.

The information-theoretic question that this chapter asks about a literary work's capacity to make this structural chain from aesthetic resonance to real-world material system subjectively intelligible is, in this sense, a question about the *aesthetics of ideology*, though in a rather different sense of 'aesthetics' from the one at play in 'new aesthetic' literary scholarship (Brian Massumi, Joseph Lavery, Ngai's later work) that studies social phenomenologies of beauty, sublimity, cuteness, exquisiteness, and so on. The more directly relevant sense of 'aesthetics' is, instead, the sense originally coined in Baumgarten's 'Aesthetica': aesthetics as the theory of sensible (*sensibilis*, sensate) cognition and of sensible representation, and in particular the theory of sensible cognitions and representations that rise to the level of sensible abstract thought (*analogon rationis*). As Ngai hints in describing ideology as a 'materially embodied representation of an imaginary relationship to a holistic complex of real conditions' and mapping its 'virtual, diffused, immanent' consistency to the idea of 'tone,' an 'ideology' in the contemporary (broadly Western Marxist) critical-theory sense is an epistemically and semiotically unwieldy object, as elusive from the point of view of analytic, literal cognition and representation as a 'tone' would be. The proposition that some literary works are functionally analogous to the canon of a trained autoencoder, then, is in part an account of our epistemic and semiotic access to phenomena of the type 'virtual, diffused, and imma-

nent, materially embodied, representation of an abstract, general imaginary relationship to a holistic complex of real conditions' in general, and to 'ideology' in the critical-theory sense in particular.

2.1 Universalia in re

The central proposition of this section is, informally, 'it's possible to learn a way of seeing by examining a group of objects that this way of seeing sees the best.' I take my cue from how unsupervised representation-learning neural nets such as autoencoders not only learn a low-dimensional representation-space (a feature function, which we called a trained autoencoder's *worldview*) but also a projection from each point in input-space to some point in a low-dimensional manifold *within the input-space*. Typically, deep learning researchers are mostly interested in this projection function (which we called a trained autoencoder's *method of mimesis*) in the training stage, where it's used as a proxy for the low-dimensional representation-space: they employ the distance between input and projection on the training data as an optimization goal, and when the training's finished they extract the low-dimensional representation-space this training has effected. Speaking as literary theorists, however, we have special reason to take interest in the lower-dimensional submanifold in input-space—that is, the set of input-space points the trained neural net learned to project into, which we have called a trained autoencoder's *canon* of privileged objects, or more formally *the image of a trained autoencoder's projection function*. The canon of a trained autoencoder, we concluded at the end of the previous chapter, effectively mathematically expresses the autoencoder's *feature function and decoder*, and so we can understand the canon to express the trained autoencoder's overall systemic logic. For those of us concerned primarily with cognitive systems that cannot directly 'download' or 'upload' computational modules

to and from each other at high bandwidth—a basic fact of life for the cognitive systems of the human world, but not for digital cognitive systems built by machine-learning engineers—the image of a cognitive system's projection function stands out from other mathematical expressions of a system for comprising individually intersubjectively meaningful parts that cumulatively express the system's own representation-space or 'subjectivity.'

While the representation-space of any trained autoencoder is composed of predications in a sui generis 'mental language' bound to the autoencoder's feature function, the canon of a trained autoencoder is, in an important sense, intrinsically ***intersubjective***. To the extent that we conceive of the representation-space of an autoencoder not just as (e.g.) the set of all lists of 100 numbers ranging from 0 to 9,[24] but as a space of complex abstract predications given by the trained autoencoder's feature function, the abstract entities that make up a given trained autoencoder's representation-space are foreclosed to any subject who does not already have a working understanding of that trained autoencoder's feature function. It's not a simple accident of our cognitive biology, in this sense, that as human subjects we don't have an abstract 'language' of representation-spaces, nor necessarily have the capacity to form clear and distinct ideas corresponding to each entity in each representation-space we make or use. A 'language' that refers directly to coordinates in one's representation-space is useless, or even meaningless, for agents whose representation-spaces are both heterogeneous (different between agents) and dynamic (different within-agent over time), and humans are almost above all else a *dialectical* animal: rarely of one mind with each other or with one's own self for very long but also ceaselessly engaged in inter-agent syncretism and appropriation. It might begin to follow, consequently, that by the same token humans should also

[24]We could also define a trained autoencoder's representation-space as simply the image of its feature function, comprising simply the set of all lists of x numbers ranging m to n. By this definition, the entities of the representation-space are concrete, but both individually and collectively meaningless without the feature function.

be a *hermeneutic* and *aesthetically expressive* animal. Outside of the digital realm, one cannot hope to transmit entities from a given representation-space to a cognitive system that has not already mastered this representation-space unless one first converts these entities into semiotic, sensory, or material entities in an intersubjective input-space. Because the entities of a representation-space are inextricable from their holistic context, however, any such 'intersubjective correlative'—or, to be less coy, 'objective correlative,' since we use 'intersubjective' here as a synonym for certain meanings of 'objective'—can only transmit a specific abstract predication within the representation-space by taking part in an 'intersubjective correlative' to the representation-space in its entirety. In other words, to make the thoughts formed in a given subjectivity accessible, one has to give the mood, gestalt, cognitive map and so on is (to borrow Heidegger's pith) 'the precondition for, and medium of' these thoughts' intersubjective form.

What we require, then, if we're to integrate the mathematical trope of representation-spaces into the world of semiotics and aesthetics, hermeneutics and communication, is a model of 'intersubjective correlatives' for representation-spaces. As the attentive reader might already guess, our paradigm comes from the mathematical trope we have called the canon of a trained autoencoder. Unlike the inextricable abstracta that make up a trained autoencoder's representation-space, the contents of the canon of a trained autoencoder are individually accessible to our sensory or semiotic faculties, as well as to the sensory or semiotic faculties of any other subject, animal or mineral, that deals with the same input-space. Nevertheless, when taken as a whole this canon of concrete, intersubjective input-objects expresses the representation-space in all its inextricable abstractness—a transparent lamp about a spiritual flame, if you will. And while the canon of a trained autoencoder is, perhaps, not quite intersubjectively concrete *itself*, since it is strictly speaking an infinite set, a sample

from the canon of a trained autoencoder faces no such barriers to concrete intersubjectivity. To the extent that an autoencoder's canon is amenable to sampling and approximation, there can be genuine, practicable 'intersubjective correlative' for a trained autoencoder's subjectivity. Generalizing some of our observations about trained autoencoders and their canons to a broader class of systems, I propose that by producing samples from the image of its own projection function, a cognitive system can communicate its own representation-space to other subjects, opening itself to hermeneutic and aesthetic inquests.

Speaking informally, the generalized model we will be constructing is a formal analogue to a kind of Viconian ('The Discovery of the True Homer') idea of art. A subject, whether animal or mineral, apprehends the world in a particular subjective way determined by the logic of her formative environment. She produces a mimesis of the world that's simpler than the world in a specific way that bears the mark of her particular subjective apprehension of the world. Subjects who did not share her formative environment can learn her apprehension of the world by trying to very exactingly apprehend the artefactual world she produced. (Importantly, as I discuss below, the analogues here are a *single* literary artwork and a set of input-space points.) I argue that this similarity is not an accident, or even per se an analogy, and the formal desiderata that our model satisfies in fact approximate a broad functional definition of an 'aesthetic symbol' in the romantic, Symbolist, Modernist sense: a sensible representation of an abstract worldview. As our argument progresses, we will introduce the bold proposal that to the extent that an actual literary work holistically constitutes a sensible representation of an abstract worldview, one major aspect of the ***meaning*** of the literary work is the manifold of some (ideal) trained autoencoder.

While one might worry, at this point, that the above proposal will reserve the concept of aesthetic meaning, or the heritage of the romantic theory of the symbol, to straightforwardly

mimetic literary practices, we will in fact find that it thrusts us in an opposite direction. Working with a definition of a trained autoencoder's canon as its set of 'zero-error inputs,' we draw a near-equivalence between curating a selection of 'near zero-error' worldly objects on the one hand and mimetically representing/interpreting a world on the other hand, as two ways to produce a sample from a trained autoencoder's canon. This near-equivalence, we argue, lends itself to a rich philosophical account of curation, installation, collage, and other (typically Modernist) aesthetic practices that deal in 'presentation' rather than representation, as well as literary practices more indirectly guided by the materiality of language or the immanence of the imagination rather than mimesis, as nevertheless 'world-directed' or mimetic in the fundamental, Auerbachian sense of constituting a subjective model of the structure of reality. Not to appropriate any existing literary-theoretic concepts, or imply conceptual analysis of any ordinary-language concept, we will call this mode of aesthetic meaning 'ambient meaning'—partially in homage to Tan Lin's literary-practical concept of 'ambient stylistics'[25]—since it deals in sensible representation of long-run, diffused, systemic structures.

By way of one final preamble, it's important to observe that the above does not mean we aim to regard the author of a literary work as, herself, analogous to a trained autoencoder, or to understand the imaginative contents of a literary work as samples from the canon of an artist's own workaday, general- purpose representation-space. Our intention is to understand the contents of a literary work as samples from the canon of the *literary work's own* trained autoencoder—as samples from the set of concrete objects corresponding to the representation-space that comes into being in the composition of the work.[26] One major

[25]See Lin, 'Disco as Operating System.'

[26]The above is not to say that it can never be the case that a literary work's representation-space is directly the expression of a worldview that is already an available mood for the author prior to the act of writing, but merely that it doesn't have to be the case.

reason that the work, and not the author, is our living trained autoencoder—in addition to standard considerations of the Latour, Barthes, and Eliot variety—is that even in as much as a trained autoencoder may be just the same thing as a mood, or a gestalt, or an ideology, or a cognitive mapping, and even in as much as a mood, a gestalt, an ideology, or a cognitive mapping is a subjectivity, we'd hardly ever want to identify an entire person with one 'subjectivity': persons live with, and in, and through, and in-between, and around structures of subjectivity like moods, gestalts, ideologies, and cognitive mappings, but trying to get too specific about the ontology of this cohabitation in any one instance is rarely a good idea. Concluding our preparations and preliminaries on this rather quietist note, we can now begin constructing our model of 'intersubjective correlatives' for representation-spaces in earnest.

The engine of this chapter's model of 'intersubjective correlatives' for representation-spaces is a principle that is both more or less implicitly accepted in the lion's share of literary theory, and rather idiosyncratic in explicit form:[27] *it's possible to learn a way of seeing by examining a group of objects that this way of seeing sees the best.* While we already know the canon of a trained autoencoder is by definition the set of all objects that a given trained autoencoder can encode with zero reconstruction error—objects that the trained autoencoder 'sees' the best—articulating the full reason our principle should elevate this particular technical corollary of the definition of a trained autoencoder's canon is the work of this entire section. Semi-formally, we might express the crux of our argument as a sequence of four propositions:

1. The canon of a trained autoencoder is the set of all the inputs that the trained au-

[27]The clearest antecedents to this formulation might be found in the critical work of Tan Lin and of Hugh Kenner. Shelley and Vico arguably approach some related formulations, and Benjamin arguably practices a strongly related critical method.

toencoder can encode with zero reconstruction error.

2. If the autoencoder's training was successful, the objects in the canon collectively exemplify an idealization of the objects of some worldly domain.

3. An untrained autoencoder x training on inputs from a trained autoencoder y's canon will very quickly converge with the trained autoencoder y.[28]

4. 1-3 generalize to a flexible class of cognitive systems.

We start by focusing our attention on the matter of a trained autoencoder's *reconstruction error* on its own worldly domain, which we have sometimes bracketed in the previous chapter with the promise of a later star turn. As we discussed in the previous chapter, a trained autoencoder algorithm constructs objects using just a handful[29] of 'respects of variation' (formally called *features*), also known to us as the *dimensions* of its low-dimensional representation-space. While it's the goal of an autoencoder's training process to produce a set of features that can accurately generate each object in the autoencoder's training set, the training set itself is, as a rule, too heterogeneous to cover with a handful of respects of variation. What the autoencoder learns, instead, is a system of features that can generate *approximate* reconstruction of the objects of the training set. In fact, the difference between an object in the training set and its reconstruction—mathematically, the trained autoencoder's *reconstruction error* on the object—demonstrates what we might think of, rather literally, as the excess of material reality over the gestalt-systemic logic of autoencoding. The *canon* of a trained autoencoder, from this point of view, is the set of all the objects that

[28]Property 3 is the product of original, if elementary, mathematical observations developed with the aid of Tomer Schlank of the Einstein Institute of Mathematics.

[29]Typical digitally implemented autoencoder algorithms have features numbering in the high dozens or low hundreds. 'Handful' is, naturally, relative and context-dependent—the crucial thing is that the number of features must be low enough to deter the network from 'memorizing' the training set.

a given trained autoencoder, its imaginative powers bound as they are to the span of just a handful of 'respects of variation,' can imagine or conceive of whole, without approximation or simplification.

This above technical theme, I would propose, should be in part familiar to us from modern and contemporary literary theory. The 'hermeneutics of suspicion,' if there was a hermeneutics of suspicion, was a data-mining process that infers from what is found and not found in the world constructed by a literary text an organon (system of cognitive schemata) that makes certain phenomena unthinkable, invisible, foreclosed to the order of things. The critic would infer, from observation of the literary work's selection of phenomena, a generative model of the work, finding the repressed or marginalized of the text within 'gaps' in the generative model: states of the lifeoworld that the generative model cannot generate. Pushing the process even further, an ambitious critic would go on to try to characterize *dimensions*—ways states of the world can be meaningfully different from each other—missing from the generative model. Contemporary cultural-materialist or ideology-sensitive readings are, as Rita Felksi argues in 'After Suspicion,' for the most part 'post-suspicion': recent social-theoretic literary critics, especially those associated with the field of affect studies, tend to differ from their predecessors by assigning reflexivity and agency to literary texts, regarding the texts as active facilitators of a critical comparison between model and reality. This modern turn places the framework of some recent social-theoretic readers—in particular Jonathan Flatley and Sianne Ngai—in a close alliance with our theoretical concerns. Specifically, this dissertation will interpret Ngai's landmark argument in 'Ugly Feelings' that a work of literature can, through tone, represent a subject's ideology—and so both represent a structure of her subjectivity and touch upon the structure of the social-material conditions structuring her subjectivity—as strongly concordant with the following proposition: a sys-

tem of 'respects of variation,' which we might define by the excess material reality that it marginalizes ('ideology'), is equivalently defined by the aesthetic unity ('tone') of whatever contingent, hypothetical, or artificial material realities it can acess in full. The canon of a trained autoencoder, we're proposing, recapitulates the ideology of a system of 'respects of variation' as a tone.

In the pages that follow, we will start to think about a trained autoencoder's features or 'respects of variation'—the structure underlying the 'imagination' of a given trained autoencoder—as a kind of constrained language or compact vocabulary. This heuristic, while potentially dangerous, is nevertheless powerful if we keep in mind just what about the generative powers of a trained autoencoder is 'like' a vocabulary. The features of a trained autoencoder, we took pains to propose in the previous chapter, work very much like a fixed list of predicates with room to write-in (e.g.) 'not' or 'somewhat' or 'solidly' or 'extremely' next to each.[30] In the original discussion, we conceived of a trained autoencoder's features as essentially *descriptive* predicates, focusing on their role in the *feature function*, the process that 'summarizes' input objects. The features of a trained autoencoder take a rather different meaning if, instead, we center our thinking on the *decoder function*—the function that turns 'summaries' (formally *feature-values*, aka *representation-space coordinates*) into reconstructions. From a viewpoint that centers the decoder function, a given list of feature-values is not a 'summary' that could apply to any number of closely related objects, but rather the (so to speak) DNA of a specific object. A given trained autoencoder's features or 'respects of variation' are, from this perspective, akin to a list of *imperative* predicates, structural techniques or principles to be applied by the constructor. The 'generative formulae' for objects

[30]More formally, we proposed to understand the features of a trained autoencoder as analogous to a fixed list of predicates with room to write-in a numerical grade from 0 to 9 next to each, where 0 means 'not at all' and 9 means 'extremely.'

in a trained autoencoder's canon are, in this sense, a list of activation values that determines how intensely the construction process (the *decoder function*) applies each of these structural techniques or principles. It is a fundamental property of any trained autoencoder's canon, then, that all the objects in the canon *align with a limited generative vocabulary.* The objects that make up the trained autoencoder's actual worldly domain, by implication, roughly align or *approximately align* with that same limited generative vocabulary. These structural relations of alignment, we will soon propose, may have a strong relationship to certain concepts of aesthetic unity that commonly imply a unity of generative logic, as in both the intuitive and literary theoretic concepts of a 'style,' or the informal aesthetic-affective concept of a 'vibe.'[31]

The fact that the canon of a trained autoencoder aligns with a ***limited*** generative vocabulary is, perhaps, the crucial leverage point that allows us to hypostatize the concept of a trained autoencoder's canon. To be a set that aligns with *some* logically possible generative vocabulary is hardly a 'real' structural or aesthetic property, given the infinity of logically possible generative vocabularies. To be a set that aligns with some (logically possible) *limited* generative vocabulary, on the other hand, is a robust intersubjecitve property. By way of a powerful paraphrase, we might say that it means the objects that make up a trained autoencoder's canon are *individually complex but collectively simple.* To better illustrate this concept ('individually complex but collectively simple'), let us make a brief digression and

[31]I take my concept of a 'vibe' from writer/musician Ezra Koenig, by way of Batuman's discussion of Koenig's early online writing: 'It was during my research on the workings of charm and pop music that I stumbled on Internet Vibes (internetvibes.blogspot.com/), a blog that Ezra Koenig kept in 2005–6, with the goal of categorising as many "vibes" as possible. A "rain/grey/British vibe," for example, incorporates the walk from a Barbour store (to look at wellington boots) to the Whitney Museum (to look at "some avant-garde shorts by Robert Beavers"), as well as the TV adaptation of Brideshead Revisited, the Scottish electronic duo Boards of Canada, "late 90s Radiohead/global anxiety/airports" and New Jersey. A "vibe" turns out to be something like "local colour," with a historical dimension. What gives a vibe "authenticity" is its ability to evoke—using a small number of disparate elements—a certain time, place and milieu; a certain nexus of historic, geographic and cultural forces.'

describe a type of mathematical-visual art project, typically associated with late 20^{th} century Hacker culture, known as a '64k Intro.' In the artistic-mathematical subculture known as 'demoscene,' a '64k Intro' is a lush, vast, and nuanced visual world that fits into 64 kilobytes of memory or fewer, less memory by a thousandfold than the standard memory requirements for a lush, robust, and nuanced visual world. In a 64k Intro, a hundred or so lines of code create a sensually complicated universe by, quite literally, using the esoteric affinities of surfaces with primordial Ideas. The code of a 64k Intro uses the smallest possible inventory of initial schemata to generate the most diverse concreta. The information-theoretical magic behind a 64k Intro is that, somewhat like a spatial fugue, these worlds are tapestries of interrelated self-similar patterns. From the topological level (architecture and camera movement) to the molecular level (the polygons and textures from which objects are built), everything in a 64k Intro is born of a 'family resemblance' of forms.

Remarkably—and also, perhaps, trivially—the relationship between succinct expressibility and depth of pattern that we see in 64k Intros provably holds for any informational, cognitive, or semiotic system. A deeply conceptually useful, though often technically unwieldy, measure of 'depth of pattern' used in information theory is 'Kolmogorov complexity': the Kolmogorov complexity of an object is the length of the shortest possible description (in a given semiotic system) that can fully specify it.[32] Lower Kolmogorov complexity gener-

[32]In the normal definition of Kolmogorov complexity, the 'semiotic system' in question must be Turing-complete: that is, the semiotic system in question must be capable of describing a universal Turing-machine. (Our own 'default' semiotic system—that is, the semiotic system of the human subject currently communicating with you, the reader—is of course Turing-complete, since we can think about, describe, and build universal Turing-machines.) In the coming discussion, we will lift this restriction, in order to allow us also talk about the Kolmogorov complexity of certain sets relative to more limited semiotic systems—semiotic systems like a given trained autoencoder's 'generative vocabulary' (decoder function). The purpose of this deviation is to save space we would otherwise have to devote to the fidgety technical concepts of *conditional Kolmogorov complexity* and of *upper bounds on Kolmogorov complexity*. We take this liberty because unlike most other mathematical concepts, the concept of Kolmogorov complexity does not have a preexisting one-size-fits-all fully formal definition, and always calls for a measure of customization to the purposes of a given discussion. For the sake of propriety, we will mark each instance of this 'off brand' application of the

ically means more, stronger pattern. A low Kolmogorov complexity—i.e. short minimum description length—for an object relative to a given semiotic system implies the existence of deep patterns in the object, or a close relationship between the object and the basic concepts of the semiotic system. One intuitive relationship of Kolmogorov complexity and poetry has been suggested by physicist/philologist Dmitrii Manin in graceful aphoristic form: 'A text should be considered together with its meaning, understood, e.g. as a fragment of (objective-subjective) reality described/referred to by the poet—or as changes effected by the poem in the reader's mind. A poem turns out to be a proof of low Kolmogorov complexity of this object, its own meaning/content (as a small program generating a large object). That's not to slight poetic achievement, but quite the opposite: remember that Kolmogorov complexity is uncomputable, so any proof of low Kolmogorov complexity has to be a handiwork of a mathematician or a poet. Or both. A poem is a theorem. Poerem. Theoem.' (Manin, D. Yu) We will revisit aspects of this idea, which bears a complex relationship to the main proposition of this section, later in this chapter, but for now we best turn our attention back to the applications of Kolmogorov complexity to trained autoencoders and their canons, by considering the Kolmogorov complexity of *sets* of objects.

When all the objects in a given set C have low Kolmogorov$^+$ complexity relative to a given semiotic system S, we will say the semiotic system S is a *schema* for C. If S is a given trained autoencoder's generative language (formally, *decoder function*), and C the canon of a this trained autoencoder C, for example, then S is a schema for C. Importantly, any schema S is in itself a semiotic object, and itself has a Kolmogorov complexity relative to our own present semiotic system, and so the 'real'—that is, relative to our own semiotic system—efficacy of S as a schema for an object c in C is measured by the sum of the Kolmogorov$^+$

concept of Kolmogorov complexity as 'Kolmogorov$^+$ complexity.'

complexity of c relative to S and the Kolmogorov complexity of S. Because one only needs to learn a language once to use it to create however many set sentences one wishes, though, when we consider the efficacy of S as a schema for *multiple* objects c_1, c_2, c_3 in C we don't repeatedly add the Kolmogorov complexity of S to the respective Kolmogorov$^+$ complexities of c_1, c_2, c_3 relative to S and sum up, but instead add the Kolmogorov complexity of S just once to the sum of the respective Kolmogorov$^+$ complexities of c_1, c_2, c_3 relative to S. The canon of a trained autoencoder, we suggested, comprises objects that are individually complex but collectively simple. Another way to say this is that as we consider larger and larger collections of objects from a trained autoencoder's canon C, specifying the relevant objects using our own semiotic system, we quickly reach a point whereupon the shortest path to specifying the collected objects is to first establish the trained autoencoder's generative language S, then succinctly specify the objects using S.

Let us call a semiotic system S a (pure) *compression schema* (or, in standard CS terminology, *lossless compression*) for a set C if the respective Kolmogorov complexity of each reasonably large subset of C is low compared to its size, and roughly equal to the efficacy[33] of S for said subset.[34] A trained autoencoder S, then, is not merely any schema for its canon C, but a *compression schema* for its canon C. Let us recall, now, that the purpose of a trained autoencoder's training is to achieve a low reconstruction error on the objects of the training set, and consequently—if the learning process was successful—a low reconstruction error on the objects of the training domain at large. In other words, the canon of a trained autoencoder is the product of an optimization process for producing *approximate reconstructions* of the objects of a training set randomly sampled from some worldly domain,

[33]Like Kolmogorov complexity itself, our measure of efficacy is defined such that lower is 'better.'

[34]This definition of 'compression schema,' while idiosyncratic, is designed to converge with common CS informal use.

and if the learning process was successful every object in the training domain (the source of the training set) is *approximately identical* to some object in the trained autoencoder's canon. If a semiotic system S is a compression schema for a set C, and every object in a set D has an *approximately identical* object in C, we will say that S is an ***approximate compression schema*** (or, in standard CS terminology, ***lossy compression***) for D. If D is the training domain of a successful trained autoencoder S, then, S is an *approximate compression schema* for D.[35] With the above concepts in place, we might now make a curious observation. Let S be a trained autoencoder with a canon C and training domain D. Now consider a second trained autoencoder, S^*, whose training domain is C, the canon of S. Because, unlike a training set drawn from ordinary real-world domains, a training set drawn from C is patently not too heterogeneous to reconstruct using a handful of 'respects of variation,' S^* is a lossless compression (a proper *compression schema*) of its training domain C, and C is therefore the canon of S^* as well as its training domain. And because S^* is a lossless compression of C, S^* is a *lossy compression* (an *approximate compression schema*) of D, the training domain of our original autoencoder S. S^* is, in other words, just S again: if we train an autoencoder on the canon of a trained autoencoder S, what we get is S.

Consider, now, the following intuitive hypothesis. Because the canon of a trained autoencoder is 'purely' collectively simple, whereas the real-world training domain of even a successful trained autoencoder is at best 'approximately' or 'roughly' collectively simple—because the canon has a clear tone whereas the training domain is matted, because the canon is idealized whereas the training domain is shot through with excess details of empirical reality, because the canon is structure 'all the way down' whereas the training domain's structure is emergent—a sample from C should make for a more potent training set than

[35] That a given semiotic system S is an approximate compression schema for a set D does not rule out the possibility that S is also a (perfect) compression schema for D.

a sample from D, even as both C and D 'lead' to S. We argue that the above hypothesis is correct, and in fact generalizes to a broad, flexible range of cognitive systems. By way of an intuitive example, we might think about a toddler who learns how to geometrically compress worldly things by learning to compress their geometrically idealized illustrations in a picture-book for children. Let m be the number of sunflowers, full moons, oranges, and apples that a toddler would need to contemplate in order to develop the cognitive schema of a circle, and n the number of geometrically idealized children-book illustrations of sunflowers, full moons, oranges, and apples that a toddler would need to contemplate in order to develop this same cognitive schema. Empirically, intuitively, and—with some plausible assumptions—mathematically, $n < m$. More technically, let $x(D_n)$ mean the output of an unsupervised learning algorithm x trained on n samples from dataset D. It's provable from fairly weak assumptions about D and x that for any reasonably large m, there is an $n < m$ such that with high probability $x(x(D_m)_n) = x(D_m)$. The set of pictures in the picture-book is, on this model, functionally close to a sample from the image of the projection function of the cognitive scheme of a circle, operating as an information-theoretically optimal sensible representation of the cognitive scheme of a circle. Compressing a *small* set of idealized images induces the same compression schema—lossless for the idealized images, lossy for the natural images—as compressing a larger set of natural images. In other words, while it is possible to learn a way of seeing by examining a very large number of objects that this way of seeing *sees well* (the natural images, objects of the 'original' training domain), it's also possible to learn a way of seeing by examining a much smaller number of objects that this way of seeing *sees the best*.

2.2 A Defense of Poetry

When dealing with cognitive systems that one can't 'crack open'—like the read/write-inaccessible cognitive systems of the human world, and unlike the digital systems built by machine-learning engineers—the image of a cognitive system's projection function is the most direct accessible manifestation of its representation-space, and the concrete (or, in the case of literature, imaginative) realization of this image by mimesis of the world becomes a key method for inputing this representation-space to new cognitive systems. In fact, considerations from $n < m$ suggest a method for transferring (lossy) compression schemas between agents, and a reason to believe that this method has a computational advantage over training each agent individually: assume a trained neural network w that has a (lossy) compression schema for a distribution D, and an untrained[36] neural network v that we want to train to have a (lossy) compression schema for D. Assume, now, for the sake of an analogy with humans, that the internal structure of the neural nets is read/write inaccessible. If w and v have broadly compatible architectures (i.e $w(D) \simeq v(D)$), the efficient and reliable solution is to train on samples from w's reconstructions of inputs from D, instead of training v di-

[36]Importantly, the actual cognitive-theoretic model I'm proposing deals with works of art as methods for communicating a compression schema to a learner that already has a moderately advanced compression schema to start with—i.e. a pre-trained net whose representation-space is low-dimensional relative its 'raw' input-space but high-dimensional relative to the size of D—rather than as methods for communicating a compression schema to a new autoencoder. While the general principles remain applicable, one crucial difference has to do with the general learning-theoretic rule of thumb that an effective training set must be *large* and *diverse*. On a 'new autoencoder' model, this rule would predict 'rich' Modernist works like 'Ulysses' or 'Zorns Lemma' or 'Tender Buttons' or 'Ketjak,' but disqualify 'thin' Modernist works like 'The Making of Americans' or 'The Trial' or '7CV' or 'Blood and Guts in High-school.' Things look very different, however, when we consider what a training set designed to update only the connections between higher layers of a network that already received a lot of broadly relevant training would be like: the training can now take advantage of the network's existing sophisticated similarity space to generalize from a prototypical instance of an everyday concept to the rest of this concept's (reasonably typical) instances, such that e.g. however 'The Metamorphosis' is training a reader to encode states of the Samsa family, after training the reader will fairly similarly encode states of the Zamsa family, who are like the Samsa family but Dutch and have red hair.

rectly on samples from D. In learning to compress the image of the trained net v's projection function, w learns a projection function that closely approximates (on D) that of the trained net v.

Building upon the above principle, we can construct a mathematically informed interpretation of a kind of minimal romantic, Symbolist, Modernist working assumption: that a work of literature can effectively communicate an ineffably complex holistic understanding of the real world, which we might intuitively call the work's 'aesthetic meaning.' On our mathematically informed interpretation, one sufficient condition for romantic, Symbolist, Modernist 'aesthetic meaning' is the sensible representation of an autoencoding manifold in an intersubjective input-space, which we will call 'ambient meaning.' Not to feign innocence, let us observe ahead of time that we will argue that at least *some* actually-existing works of literature in fact construct 'ambient meaning,' and the minimal romantic, Symbolist, Modernist working assumption is well-integrated with our best social-theoretic and technoscientific understanding of the realms where it must operate. Our construction of this 'mathematically informed interpretation' will, in this sense, take more after the dialectical form of a 'defense' (as in 'Defence of Poetry) or an apologetics[37] than after the dialectical form of disinterested scholarship. Where scholarship is predicated on commitment to a specific engine of reasoning and a promise of disinterest in the end-result of reasoning—the scholar goes where the unbiased application of her method leads—the stance of an apology admits a prior commitment to a certain end-result, and seeks to demonstrate that this commitment is compatible with certain prima facie challenging methodological commitment. Moreover, the inevitably theological connotations of 'apology' are especially apt here, since reification of 'romantic, Symbolist, Modernist' forms of insight, sense, or meaning that are

[37]We might consider Rita Felski's 'Uses of Literature' or Caroline Levine's 'Forms' as two contemporary examples.

not accessible to common sense or ordinary analytic thought is oftentimes associated with what Rita Felski recently and accurately nicknamed 'theological' (3) accounts of literature. In its capacity as an apology, this dissertation makes the case for literature's 'otherworldly aspects' (ibid) to the twice-materialist magisteria of social theory and technoscience: on our account, the 'otherworldly' is an aspect of the natural and social world that is 'foreclosed to 'analytical' and concept-driven styles of political or philosophical thought,' (ibid) but whose interactions with conceptual cognition and with everyday practical knowledge are abundant and empirically tractable, and whose foreclosure is itself analytically and technoscientficially explicable. While neither the form nor the substance of 'ambient meaning,' as we have it, finds a place within the order of things as revealed by commonsense materialisms, the forms and substance of ambient meaning are plausibly all but endemic to the language of a *mathematically sophisticated* technoscientific materialism, and to the language of an *affectively sophisticated* cultural materialism.

The proposition that a work of art can model the manifold structure of a real-world distribution is, perhaps, particularly pressing in its relevance to Modernist aesthetic theory and practice, in as much as Modernism is stereotypically associated with a focus on abstract, diffused, ineffable structures of meaning bound to a work's overall aesthetic form rather than on more direct or local cognitive-affective fruits of a work's narrative, rhetorical, and lyrical communication. We will regard as 'stereotypically Modernist,'[38] for instance, the full corpuses of Jarry, Woolf, Kafka, Maeterlinck, Roussel, Pound, TS Eliot, Stein, Musil, Bely, Shklovsky, Benjamin, Khlebnikov, Kharms, Mishima, Beckett, Pinter, Ashbery, Saurraute, Haroldo de Campos, or Robbe-Grillet, and of staple 'proto-Modernist' anchors like Büchner, Melville, Lautreamont or Dickinson, as well as parts of later Goethe, Charlotte Bront, later

[38]We of course use 'stereotypically' here in a value-neutral sense.

Chekhov, and later Flaubert. While it is plausible, in my view, that 'ambient meaning'—in effect, the sensible representation of a systemic gestalt—is already an endemic aspect of 'sensible discourse' as Baumgarten understood it in 1735, and that many or most literary works in many or most literary canons function in part as holistic aesthetic symbols in the relevant sense, our model nevertheless has a certain rich 'elective affinity' with Modernist, as well as 21st century 'experimental,' literary practices and literary thought. First, and most crudely, an account of a given literary work's ambient meaning plausibly explains *more* in the case of a stereotypically Modernist work than in the case of, for example, a Regency novel. It's useful, here, to draw on Roman Jakobson's conception of a literary work's '*dominant*': a work's subsuming semiotic structure, whereby a work's non-dominant semiotic structures operate as lower-order organizing principles downstream from the dominant's logic. Even rejecting, as we should, the rigidity of Jakobson's idea of a perfectly teleological, means-end relationship between a work's various semiotic structures, it's fair enough to say that from among reliably salient organizing principles like rhetoric, narrative, lyricism, verisimilitude, and holistic aesthetic symbolization, the logic of holistic aesthetic symbolization tends to the condition of Jakobsonian dominance in the semiotic economy of a stereotypically Modernist literary work. Whatever we believe about just how exhaustively e.g. 'Bouvard et Pécuchet' (or even 'Le Dictionnaire des idées reçues') or 'The Making of Americans' are optimized for the sensible representation of a trained autoencoder's manifold, there can be little doubt that 'Pride and Prejudice' comprises a wealth of aesthetically and ideologically salient semiotic operations besides 'ambient representation.' Stereotypically Modernist works, on the other hand, are often marked not only by structural subsumption, but also by zero-sum tradeoffs whereby typically salient or pressing semiotic structures like plot, agency, logos or are broken or matted to allow additional 'degrees of freedom' for the work's holistic aesthetic symbol-

ization structure. At the limit of stereotypically Modernist practice—'Finnegans Wake,' 'The Making of Americans,' "'A,"' 'Galáxias'—we are left with perhaps nothing other than holistic aesthetic symbolization to hang our hat on.

The final point above is, perhaps, particularly relevant to this dissertation's project from a practical or dialectical perspective. While—for example—Tolstoy famously asserted that the sense of 'War and Peace' is inextricable from the totality of 'War and Peace,' and while I would propose that 'ambient meaning' as per our autoencoder model is a major aspect of this total meaning in the case of 'War and Peace' perhaps no less than in the case of 'Tender Buttons,' an interlocutor may peacefully reject holistic aesthetic symbolization as nonsense upon stilts and still believe that 'War and Peace' not only delights but instructs, delivering many individually fine lessons about human nature through its piths and parables. The bare assertion that 'War and Peace' plausibly yields substantive cognitive-affective content doesn't, on its face, present us with an explanundum that calls for a mathematically augmented theory of 'ideas in things.' When dealing with a stereotypically Modernist work of literature, on the other hand, ascription of substantive cognitive-affective structure necessarily courts theoretical controversy. In fact, the question of the cognitive robustness of stereotypically Modernist meaning-making even breaks away from the glorious isolation of 'pure' or transhistorical aesthetic discourse, and powerfully interpolates our theory of 'ambient meaning' into a dialectical relationship with what we might call the default Marxian story of representation under late modernity, as we discuss below.

Marxian literary history, if there is *a* Marxian literary history, contends that under late modernity the causal structure of social-material conditions of subjective life has become too complicated to abide a narrative or lyrical representation, and modern literature has therefore been increasingly unable to provide 'cognitive mapping' of the social-material con-

ditions of subjective life. An 'ambient meaning' viewpoint on Modernist literary practices accepts the Marxian literary historian's premise while rejecting her conclusion, arguing that the conditions of modernity do not necessarily diminish literature's overall capacity to structurally model social-material conditions, but perhaps modulate the capacity of narrative or lyric tactics of cognitive mapping while amplifying the capacity of 'ambient' mapping strategies. I have hinted above that the 'default Marxian narrative' about representation under late modernity enjoys a certain pro tanto authority that is sufficient for putting ascriptions of substantive cognitive-affective modeling capacities to stereotypically Modernist literary work in the position of *apologetics*. The factors that determine our dialectical positioning here are, perhaps, a contingency of the particulars of Anglophone literary-theoretic discourse on forms,[39] but they have nevertheless found reification through one more successful illocution in 20$^{\text{th}}$ century literary academia. Speaking at the now-legendary 1983 'Marxism and the Interpretation of Culture' conference, Frederic Jameson gave a keynote lecture called 'Cognitive Mapping,' that will later become the foundation for the most heavily cited monograph in literary theory:

> 'I am addressing a subject about which I know nothing, whatsoever, save for the fact that it does not exist. [...] Since I am not even sure how to imagine the kind of art I want to propose here, let alone affirm its possibility, it may well be wondered what kind an operation this will be, to produce the concept of something we cannot imagine.' (347)

In his lecture, Jameson hypothesized a Marxian aesthetic of the future, called 'the aesthetic of cognitive mapping,' that is waiting to be born and as of yet unknowable. Declaring

[39] As Caroline Levine observes, the systematic study of aesthetic forms in late 20$^{\text{th}}$ century literary theory has been, for the most part, the domain of a single Marxian lineage.

from the start that he has nothing to say *about* this yet-unborn aesthetic, Jameson left the baptism itself to be the only anchor of the future aesthetic's identity. And like a Catholic Walter Shandy, Jameson set the fortunes of his unborn Marxian aesthetic for the next four decades with the act of giving it a name. Naming the concept after a beloved senior relative from Marxian urban planning theory—the concept of 'cognitive mapping' in the work of urban theorist Kevin Lynch, where 'cognitive mapping' denotes a city resident's ability to comprehend the spatial, social, and functional layout of her city as an integrated living system—Jameson's Marxian baptism proved at least the equal of its Roman-Catholic antecedent in illocutionary force. All future talk of an aesthetic practice that is epistemically constructive (not unlike a city resident's cognitive mapping of her city), systems-minded (not unlike a city resident's cognitive mapping of her city), and holistic (not unlike a city resident's cognitive mapping of her city) was to become, at least partly by power of this Jameson's 1983 illocution, a claim on Jameson's hypothesized 'aesthetics of cognitive mapping.' It should be obvious, then, that ascriptions of 'ambient meaning' to a literary work are in some sense a claim on Jameson's 'aesthetics of cognitive mapping.' While Jameson's lecture does steer clear, as promised, from describing what his 'unimaginable' aesthetics of cognitive mapping might be like, Jameson does forcefully argue that there are no examples of an actually-existing aesthetics of cognitive mapping—and therefore, on my terms, that there are no actually-existing literary practices of ambient meaning-making. Moreover, Jameson's 'Cognitive Mapping' lecture offers a rather thorough study of its namesake aesthetic *via negativa*, by recounting the Lukácsian story about literature's loss of powers of representation under the conditions of modernity.

On Jameson's 'Cognitive Mapping' account, orthodox to what we might call the Lukács-Jameson-Moretti school of Marxian literary theory, Modernist forms are primarily dramatic

performances of epistemological desire, not good so much for the modeling of worldly systemic structures and relations as for dramatizing (symptomatically or critically) modernity's progressive crises of representation and of social knowledge. This dissertation both accepts—in as much as every literary-theoretic enterprise operates from *some* implicit or explicit normative stance, and this is ours—the Lukács-Jameson-Moretti teleology wherein a literary art's success or failure as an epochal enterprise lies in its capacity to model systemic structures and relations (not, let's say, a literary art's capacity for stirring passions, or for immersive storytelling, or for free-play of the faculties), and rejects the Lukácsian tradition's actual diagnosis of late 19th - 21st century literature's capacity to model systemic structures and relation. By contrast with the Lukácsian tradition, the present work treats Modernist and avant-garde forms not as epistemic dramas but as epistemic methods. This commitment goes beyond objecting to the often caustic treatment of Modernist projects in the work of Lukács, Jameson, or Moretti proper, and challenges a rather flexible and porous paradigm associated with the rich diffusion of Jamesonian (Lukácsian) ideas in Modernist studies at large. Where the work of Lukács, Jameson, or Moretti proper tends to cast Modernist forms as tellingly miscalibrated readjustments of literature's epistemic project to the objectively contracting possibilities for social knowledge and representation, critics and theorists who are broadly Jamesonian (Lukácsian) but sympathetic to experimental literary practices typically defend Modernist forms as tactics of resilience, wherein literature lowers its representational ambitions in the face of the objectively contracting possibilities for social knowledge and representation in order to make the most of what can still be represented.[40] By contrast with even these (so to speak) Jameson-compatible defenses of Modernist representation, this

[40]Sianne Ngai's chapter on 'stuplimity' in 'Ugly Feelings,' a book which is otherwise an inspiration for the present work in many respects, is a prototypical example of an account of Modernist forms as tactics of *resilience*.

dissertation's point of view suggests that Modernist forms are not entirely a reconciliation of literature's epistemic ambitions to an epoch's already-configured epistemic possibilities, but also an invention or discovery of novel[41] epistemic possibilities that epochal conditions have made newly practicable, thinkable, or urgent.

From the viewpoint of Modernist studies, the theory of ambient meaning can propose new possibilities for a *concrete* constructive account of the capacities of Modernist and avant-garde textual forms for modeling social, psychical, and cultural-material structuring structures: concrete, at first, in the high specificity of its recourses to social-theoretical and cognitive-theoretical discourse when defining the semantics and epistemology of the Modernism-friendly (to say the least) form of sensible representation that I call 'ambient meaning,' and resultantly concrete in the high specificity of the relationship that it imputes between the historical and formal vicissitudes of Modernist textual practice and the historical and cognitive vicissitudes of making, sharing, using, representing, and scrutinizing mental models of the social. Modernist forms, from this viewpoint, are concrete adaptations or developments in the technology of cognitive mapping—they are pragmatically sociohistorically contingent[42] *methods* for creating mental models of the structures and dynamics of phenomena. While such a research program as described above is well beyond the purview of this present effort, in the course of our study of ambient meaning we hope to give solid ground to the hypothesis (redundant to some, hopeless to others, and therefore maybe relevant to all) that Modernist and avant-garde forms are a viable, and sometimes vital, cognitive technology

[41] Much like in the history of science, philosophy, or technology, we should read 'invention,' 'discovery,' or 'novel' so strongly as to suggest historical discontinuity or lack of antecedent.

[42] We bracket the question of whether or to what extent the relevant form of sociohistorical contingency is ideological in additional to practical. That is, it may or may not be the case that the rise of Modernism was simply a function of 'motive and opportunity' for certain kinds of cognitive-aesthetic literary operations, and it may or may not be the case that the rise of Modernism was a result of some ideological shift in the criteria of truth, knowledge, representation and so on.

for representing sensible systemic structures in a social-material.

Speaking in more literary-critical terms, we might say that the 'ambient meaning' viewpoint treats Modernist form as a *method of map-making*, rather than as a map:[43] I treat abstraction, parataxis, fragmentation, indeterminacy, polysemy and polyvalence not as representations of a psychological or cultural predicament but as methods of a process of cognition. We will read radical fragmentation in a Modernist text, for example—all else being equal—as part of the functional structure of a learning/teaching process that proceeds by way of a curated field of heterogeneous objects to impute whatever representation of the mind or the world, rather than as a representation of a fragmented mind or a fragmented world. We thus partially detach the questions of Modernist form from questions about the form of modernity, of the world-system, of our Worldedness or of Lukács totality—not in order to deny the possibility of a deep epochal relationship between the cognitive and social conditions of modernity and Modernist form, but to propose that this relationship is at least in part mediated by the extent to which the conditions of modernity are especially enabling of, or make a special necessity of, the cognitive work of producing ambient meaning. One stance which this work does take on matters of form and epoch, however, is that every degree of abstraction, parataxis, fragmentation, indeterminacy and polysemy is consistent with every orientation towards social totality, social brokenness, capitalism, fragmentation of society, affect, or what you will. In fact, I would suggest that it is something of a catastrophe of the academic Modernist canon that when we think of Modernist collage we think

[43]In 'Plurality in Question,' Jackson instructively makes a complementary, 'opposite' point in a very different context. Jackson observes that critics often read the typical rhetorical and narratological linearity and rigidity of the contemporary Zimbabwean novel as imputing a linear, rigid world or worldview, and takes issue with the reduction of the representational function of literary form to a direct analogy between the discursive structure of a novel and the structure of a social-material totality. On Jackson's account, the narratological linearity and rigidity typical of the contemporary Zimbabwean novel is instead an aspect of its *functional form* as a dialectical or argumentative novel, a novel whose structure constructs a particular thesis rather than simulates a totality.

only of Eliot's 'heap of broken images' and not of Zukofsky's 'integral, lower limit speech, upper limit music,' or that when we think of parataxis we think of Tristan Tzara's histrionic nihilism but not of William Carlos Williams's earnest repleteness.

There is a visible relation, one might notice, between our gloss on 'stereotypically Modernist' literature and what Marjorie Perloff's 'The Poetics of Indeterminacy' famously called 'the 'other' tradition'—a subtradition within Modernism that Perloff identifies (in poetry) with Rimbaud, Pound, Williams, Stein, Ashbery, Beckett, and Cage, and against Baudelaire, Mallarmé, Eliot, Stevens, Crane, and Lowell. Perloff's 'other' tradition is a useful marker for the Modernism I seek to associate with the production of ambient meaning, but more useful yet is Perloff's reflection, some thirty years later, that the Modernist 'other' tradition is perhaps not an exact historical class of literary corpuses, but a literary-theoretic and literary-historical *point of view*:

> 'I contrasted this "other" tradition to the TS Eliot "Symbolist" one. In retrospect, of course this dichotomy is specious. Eliot, surely a major influence on Ashbery, could just as well have been placed in the opposite camp, and of course Mallarmé could have replaced Rimbaud. I think what I really had in mind was less the poetry itself than the way it was being read. Eliot's New Critical heirs were read for their "symbolic" meanings and you couldn't do that with Ashbery. So from that point of view, the distinction was, and I hope still is, useful.'

In fact, the difference was perhaps as much in who was reading Eliot and Crane and who was reading Pound and Ashbery—and, circa 2006, started developing a taste for Eliot. Perloff's 'Poetics of Indeterminacy' was the work of a liaison between late New Critical, early post-structuralist academic poetry criticism and its own 'other tradition,' the commu-

nity of para-academic literary scholars and practitioners that grew out of the Black Mountain school, Cage's circle, and the New York School in the 60's, began to consolidate its counter-institutions and counter-canon in 1971 with 'THIS' press and its companion magazine 'THIS,' and by 1981 already founded the iconic 'L=A=N=G=U=A=G=E' journal. Still, it is more than reasonable to ask what kind of practice takes the Poundian collage of heterogeneous, barely-networked, but richly concrete nodes—concrete as texts, or concrete as images, or concrete as concepts, and often concrete in all three regards—together with the small, molecularly unstable Ashbery lyric, but not the Baudelairean symbol that ponders 'secret and intimate connections between things, correspondences and analogies.' In what sense, or through what kind of reading, could an Ashbery lyric, let alone a Rimbaud lyric, be more Poundian than Baudelerian?

With apologies to my reader, I reserve my own treatment of an Ashbery lyric ('At North Farm'), which well serves as a first experiment in 'ambient reading,' for the final section of this dissertation. Perloff, writing in 1981 under the auspices of 'indeterminacy,' finds a welcoming junction between Poundian concreteness and Rimbaudian instability in the then-popular trope of difference/deferral, understood as a locus of both the materiality of language and the instability of meaning. From the perspective that I favor, this theoretical instinct for analogizing Ashbery's or Rimbaud's practice and a Poundian field-composition at the level of the signifier is misguided in a particularly instructive way. It's natural enough, in this respect, to map the unstable relationship between semiotic parts in a Rimbaudian poetic practice to the paratactic freedom of semiotic parts in a Poundian field composition, but the aesthetic of entangled semiotic parts in a disequilibrium is not, or at least not for every kind of reader, in great aesthetic sympathy with the vistas of Pound, Olson, or Zukofsky field composition. An aporia of semiotic entanglements and subsumptions as in Ashbery or

Rimbaud is, to the contrary, a proliferation of relations drowning the relata—the *things*—out. In fact, I would propose that if there is a genuine shibboleth for the 'other' literary-critical and literary-practical community as it continues to this day, it's a certain silent rejection of the prima facie Modernist or radical appeal of aporia, by contrast with e.g. Paul de Man's celebratory deconstructionist readings of English romantic verse, or the literary-practical mainstay of post-Modernist fiction. (We may recall, for instance, that Louis Zukofsky, the immediate Modernist model for Language Writing, describes poetry as 'the detail, not mirage, of seeing, of thinking with the things as they exist, and of directing them along a line of melody,' or that Ron Silliman described 'Ketjak' as one 106 page long syllogism.) In fact, from Perloff's later 'Radical Artifice,' to Barrett Watten's 'The Constructivist Moment,' to Sianne Ngai's use of 'constructivist' to describe the unifying genre of the literary lineage that runs through 'Ugly Feelings,' the idea that—far from being deconstructionist in even the philosophically subtle sense—'constructivism' is the tradition's best name repeatedly, more or less independently asserts itself. As such, the term 'constructivist' operates in this context not so much as a shared literary-theoretic concept, but as an occasion for improvisation. Watten, for example, explicitly constructs his concept of 'constructivist' as a massively overdetermined polyseme, referring to historical Russian Constructivism, to the foregrounding of a work of art's formal construction, to the society-making work of utopian vision, to Freud's 'Construction in Analysis,' and to the social construction of subjectivity.[44] Ngai, describing the typical formal profile of the works that appear in 'Ugly Feelings'—a study not officially about Modernism at all, but one whose literary-historical outlook, and implicit literary canon, is drastically shaped by Ngai's decade in post-Language literary practice—negotiates her choice by calling the selected works 'constructivist,' (10) in quotes.

[44]See Watten xix.

The Modernist literature we take to have an 'ambient meaning' *dominant*, then, is a 'constructivist' literature.[45] What 'constructivist' will mean, on this occasion, is that it is a cognitive-aesthetically constructive literature—that is, a literature that deals in learning more than in unlearning—and that it is a literature that presents more as a kind of construct, a kind of structure or a system or machine, than as a speech act. For a 'constructivist' literature, uninterpretability is a basic condition for the artwork, not a paradox or a reversal or deferral of resolution. To the degree that this requirement of non-interpretation is motivated by concerns about the capacity of interpretive closure to suppress richer or wilder shades of meaning, its game is not to defer the reduction of hermeneutic engagement to interpretive resolution, but to ***obviate*** the reduction of cognitive-aesthetic engagement to hermeneutic engagement. What does this mean? Imagine, for example, that somebody asked you what 'The Trial' proves about the world. Not what 'The Trial' *says* about the world—we have interpretation for that—but what 'The Trial' actually proves. A strange question, surely, even if a little similar to ones we've heard every few hundred years since Aristotle. We might answer that the question seems askew, that the relationship of literature to knowledge goes through testimony, trust, communication. (Or perhaps we answer in our critical-theoretical voice, talking instead of author-functions, power, and performances of knowledge.) We clarify, in either case, that to say that a literary text produces knowledge is to *take it on its word*, to trust that what the text communicates to us expresses knowledge. Still, our interlocutor insists that what she wants to know is what is it that 'The Trial' would teach you equally whether she told you it's a book by Kafka or the product of a thousand monkeys on a thousand typewriters—what is it that you'd learn if you were visiting the Library of Babel and

[45] In calling a work of literature 'constructivist,' or—as we do later—'radically aesthetic,' I mean primarily to imply that that a strongly constructivist viewpoint is congenial to the work. I do not mean to imply that a constructivist viewpoint is the only viewpoint congenial to the given work. Let a thousand partially overlapping literary-historical classes and literary-critical paradigms bloom.

picked up 'The Trial' from one of the endless rows of shelves. Curiouser and curiouser, then: the royal road from literary text to knowledge goes through testimony, and testimony in turn goes through interpretation—through an author-function that says what the work believes and lends it epistemic weight. To ask what 'The Trial' proves—what knowledge it produces outside of the worth of testimony—is to ask what's left of literature as an epistemic practice if we take away interpretation. What **can** a work of literature prove, before interpretation turns it into testimony? Nothing in particular, we'd have to say. This dissertation's point of view aligns with that part of Modernist literary practice we might understand as working to make everything out of this nothing—as working to construct the epistemic engine of a work of art on the underside of interpretation.

This dissertation's project, then, is partly aiming to begin a story about the epistemic productivity of Modernist texts that is radically aesthetic in a Kantian sense (in that it regards what depends on testimony and interpretation as external to the work of the Modernist text), passably materialist in the technoscientific sense, and at least passably materialistic in the Marxian sense. With these broader pursuits in mind, I propose the framework of ambient meaning in part to propose a treatment of Modernist and avant-garde texts as both uninterpretable and cognitively constructive. The driving force behind this dyad comes in part from the ethos of contemporary 'ambient writing' practice—what is sometimes and to no one's benefit called 'post-conceptual poetry'—in the United States: anti-referential, anti-metaphorical, anti-Idealist, and anti-expressionist, but nevertheless conceiving of itself as strongly ideas-driven and world-directed. Developing in large part out of Tan Lin's 'Ambient Stylistics,' 'Heath,' and '7CV,' the uninterpretable practice of young experimental writers like Sophia Le Fraga, Gordon Faylor, Trisha Low, or Gabriel Ojeda-Sague positions itself not as a body-without-organs or a trans-sense, but as a cognitively constructive practice

(arguably closely related to the genre Northrop Fry calls an 'anatomy') whose productivity does not operate via interpretability—a Steinian sort of Symbolist practice, we might say.[46] My approach to canonical Modernist and avant-garde texts as works of 'ambient meaning' is, in this sense, partly by way of proposing a particular 21st century story about what the Modernist enterprise in literature meant. I believe that this 'autoencoding' story is, perhaps, our latest best hope (if one is inclined to hope this way) for a culturally[47] relevant aesthetic story about radical Modernist form. Where the heroic age of 20th century literary theory celebrated radical aesthetics in the name of the erotic[48] others of interpretation—semiosis, the body-without-organs, jouissance, the being of language, 'the third meaning'—the backlash of a more Bourdieusian age leaves little room for the radically aesthetic, or for anti-interpretive critical practice, if in the contrast between interpretation and aesthetics the aesthetic is configured as autonomous, immanent, anti-epistemic. Nor does the radically aesthetic fair much better, from this angle, when configured as endlessly differencing and deferring, self-negating, and again anti-epistemic.

To think about Modernism through the prism of its epistemic engine, I would argue, is to reconsider the radically aesthetic by taking 'aesthetics' back to its older Baumgratenian sense as 'sensible knowledge,' and taking 'radical' back to its older sense of 'root.' From this 'radically cognitive-aesthetic' point of view, interpretation fails to access the heart of Modernist and avant-garde practices not because interpretation treats the text *too much like an epistemic enterprise*, but because interpretation doesn't treat the text like *enough of an epistemic enterprise*. Interpretation asks 'what does the text believe' but not 'what does the text prove, demonstrate, establish.' As philosopher of science Kenny Easwaran has argued in

[46] See for example Trisha Low in Bomb Magazine.

[47] 'Cultural relevance' of the—admittedly rather vague—kind at question above may, or may not, be political or social relevance 'in the last instance.'

[48] See Sontag.

discussing the epistemology of mathematical proofs,[49] in a sufficiently epistemically rigorous (in the relevant sense) cognitive engagement with a text, the very concepts of intention and interpretation lose their grip on the reader's engagement:

> 'If someone presents a sequence of propositions for my consideration, and each proposition is such that mere consideration of it in light of my current beliefs leads me to believe it, then I can learn quite a bit from this person, even if I do not trust him. For instance, if he presents a deductive proof of some conclusion, I do not have to believe anything he says, as long as I independently have a high credence in the premises, and see independently that each step follows from previous ones. (...) This is related to some counterexamples to Grice's thesis about speaker meaning. Grice claimed that for a speaker to mean something, she must intend that her intentions play a role in generating a belief in the listener. However, with a deductive proof from shared premises, the intentions are irrelevant. This seems to be the case with many sorts of arguments beyond strict deductive proofs from shared premises.'

Interpretation, as opposed to 'mere' attention, association, and reflection, matters epistemically exactly to the extent that we stake the epistemic productivity of our engagement with a work on the work's value as *testimony*.[50] Or, in a more critical-theory parlance, interpretation determines what world-directed propositions the author or author-function puts

[49] Easwaran, private correspondence (2017): 'I like to summarize my transferability thesis as 'show, don't tell,' so it makes sense that the original use of that phrase exemplifies it in a sense though here it's about a way of seeing or a conceptual category or something like that, rather than merely the world itself.' We might observe, incidentally, that 'show, don't tell' found its way into MFA culture by way of Paul Engle's interest in Pound.

[50] In speaking of 'testimony' above, we mean to refer not to the direct assertoric content of e.g. an author's narration, but to the ideological (in the broadest sense) content of the author's total semiotic act—what junior-high English class would call an author's 'message.'

her epistemic weight behind. The Foucauldian 'author function' is, in this sense, equally foundational to the epistemology of traditional literary reading or (most versions of) New-Critical close reading and to the epistemology of 'subversive' reading wherein the authorial testimonial speech-act is put to a cross-examination or a 'hermeneutics of suspicion,' but external to the epistemology of radically cognitive-aesthetic reading. 'In mathematics,' Foucault writes, 'reference to the author is barely anything any longer,' and in fact we are free to push our analogy to mathematical proof even further. From the viewpoint of practical[51] mathematical formalism (e.g. David Hilbert), a mathematical proof does not **assert** a mathematical truth by means of its propositional content, but instead materially **demonstrates** a logical-mathematical truth about its own semiotic system[52] by its own structure, which we then extend through structural analogy to other worldly systems or to Plato's heaven as we wish. From the viewpoint of radically cognitive-aesthetic reading, a literary work materially demonstrates an information-theoretic truth about its own semiotic self by its own structure, which we then extend through metonymy of content to other, worldly systems or to Plato's heaven if we wish. In the next chapter, we set forth to characterize this potential relationship of a literary work's semiotic structure to other, worldly systems and to Plato's heaven both.

[51]Practical mathematical formalism, as opposed to philosophical mathematical formalism, is a viewpoint on mathematical practice rather than a theory of its ultimate ontological and epistemological foundation.

[52]We can define the 'semiotic system' as the Hintikka set model of ZFC. See Smullyan 209.

Chapter 3

Ambient Meaning

I begin by arguing that we may treat the content of a literary work as, in part,[53] a set of data-points within the reader's prior input-space. The kind of pattern-recognition practices typical of literary or aesthetic reading, I propose, consistently call on the reader to discretize the work's content into intercomparable data-points in order to appreciate the play of difference, repetition, and variation that makes up a literary work's narrative and rhetorical progression: to attend to a literary work aesthetically is, in part, to compare and contrast the various situations, objects, actions, places, characters, tropes, concepts, and images that vary over the work's narrative and rhetorical course. The concrete imaginative content of a literary work is therefore, at some practicable level of abstraction, a sequence of intercomparable data-points. For the purposes of the current thesis, we put matters of sequencing aside[54] and focus on the set of intercomparable data-points associated with a given literary work. The motivation for this reduction has less to do with the technical or conceptual limitations of our general mathematically informed framework than with the limitations of this dissertation's dialectical scope: in the long run, we can integrate matters of sequencing directly into our

[53] As I proposed above, the operation here discussed is perhaps especially relevant to stereotypically 'constructivist' or 'radically aesthetic' Modernist works of literature.

[54] We are bracketing both sequencing in (to use the Russian formalist term) the sujet and sequencing in the fabula.

present technical/conceptual model of literary 'ambient meaning' by adding considerations of 'curriculum learning,' the gradual development of a model-in-training as it takes new data-points, into our analysis, or by identifying the work's data-points with stages of a growing 'everything so far' sequence as in a recurrent network, or by encoding data-points using predictive 'skip' representations that treat an object as a Peircean index to the objects that precede and follow it. The *sets-based* point of view, however, is both uniquely conceptually transparent, compared to more technically intricate architectures involved in unsupervised representation learning on *sequences*, and attractively biased towards capturing exactly those aspects of a literary work's structural and aesthetic logic that are most orthogonal to the relatively well-understood logics of narrative, lyrical, and rhetorical structures of meaning.

One primary structure of zero-interpretation meaning in a literary work, I argue, is the collective aesthetic unity of all the objects or phenomena directly spanned by the fictional or notional world that the work imaginatively constructs. The meaning of a literary work like Dante's 'Inferno,' Beckett's 'Waiting for Godot,' or Kafka's 'The Trial,' we would like to say, lies at least partly in an aesthetic 'vibe' or 'style' that we can sense when we consider all the myriad objects that make up the imaginative landscape of the work as a kind of curated set. On the account that I'm proposing, a literary work like 'Godot' or 'The Trial' has an 'ambient meaning' to the extent that the compression schema (i.e. trained autoencoder) that solves locally linear unsupervised learning on the set of data-points that make up the work is also a compression schema for some real-life domain D. More formally, we might state this as follows. Let x('The Trial') mean the output of an unsupervised learning algorithm x trained on the data-points that make up the concrete imaginative content of 'The Trial.' To the extent that x('The Trial') $\simeq x(D_m)$ for some non-trivial real world domain D such that

$m > |$'The Trial'$|$,[55] I argue, the compression schema optimal for the data-points of 'The Trial' is a robust form of 'ineffably complex holistic understanding of the real world,' and apt to play the role of 'ambient meaning.' A little more precisely, I identify the 'ambient meaning' of a literary work with the *input-space manifold* equivalent to a compression schema (i.e. trained autoencoder) of the form discussed above. The several computational uses of this manifold, I propose, each mirror an important literary-theoretic view of the cognitive-affective functions of 'ambient meaning,' and give a wealth of nuance to the relevant notion of an 'ineffably complex holistic understanding of the real world.'

The form of 'ambient meaning,' we have argued, is an autoencoding manifold in an intersubjective input-space. But what is, so to speak, the meaning of an ambient meaning? In her monograph 'Ugly Feelings,' Sianne Ngai describes a useful commonality between certain expansive concepts of a literary work's 'tone' we find in formalist and aesthetic literary criticism, certain expansive concepts of a phenomenal world's 'mood' we find in phenomenology and philosophical psychology, and certain expansive concepts of a culture or society's 'ideology' we find in Western Marxism and in critical theory. In all these cases, Ngai observes, we identify a certain world (e.g. a certain literary text's imaginative universe, a certain person's, or a certain social ecology) as the material or sensate form of a subjective stance (e.g. a certain 'tone,' a certain 'mood,' a certain 'ideology') toward reality at large. Inspired in part by Heidegger's philosophy of 'mood,' Ngai analyzes these diverse concepts of material-subjective structures as each similarly 'virtual, diffused, and immanent' (Ngai). While Ngai coins the description 'virtual, diffused, and immanent' only in passing, I'd argue that it's useful to formally posit 'virtual, diffused, immanent structure' as the ontological category encompassing structures like tone, mode, and ideology.

[55] '$|X|$,' by common mathematical notation, stands for 'the size of the set X.'

What does this concept of a 'virtual, diffused, immanent' structure mean? We can begin with the idea that tone, mood, and ideology are 'virtual' phenomena. The 'virtual,' in Ngai's loosely Deleuzian dialect, is the hermeneutic or semantic layer of reality: the edifices of significance, meaning, and sense borne by the scaffolding of a material world. While the Deleuzian domain of 'virtual' phenomena overlaps at least roughly with a kind of Kantian 'supersensible'—the realm of epistemic, ethical, teleological, modal, affective, aesthetic, and semantic judgments grounded in the application of a transcendental schema to a particular material reality—it is a crucial feature of Deleuzian thought ('transcendental empiricism') that each material reality immanently generates its virtual phenomena, by aesthetically structuring a transcendental schema adequate to their production.[56] To say that tone, mood, or ideology are 'virtual' properties of material or sensate worlds, in this sense, is roughly to say that a world's tone, mood, or ideology is a kind of transcendental subjectivity immanent to this world's material structure. Within the context of Ngai's discussion of tone, mood, and ideology, a 'virtual' property of a sensate or material world is what we might call a 'world-embodied worldview'—that is, a tone, a mood, or ideology is a worldview embodied by a world. We might consider, for example, the pastoral mood embodied in a meadowland, or the 'Kafkaesque' tone embodied in the imaginary world constructed by Kafka's 'The Trial,' or even the commodity-fetishist ideology embodied in a capitalist market. Each structure of the 'virtual, diffused, immanent' kind, I argue—structures like 'pastoral mood,' 'Kafkaesque tone,' 'commodity-fetishist ideology'—is at one and the same the material structure of a certain constrained *world*, the schema of a *way of seeing* or a *worldview* on reality at large, and an affective quality of the world-structure and the worldview both. Moreover the 'world structure,' in each case, is at first the material

[56]See Roffe.

structure of e.g. a meadowland, or of the fictional world of 'The Trial,' or of a capitalist market, but also in some sense an aspect or a part of the material structure of whatever world—respectively perhaps a pastoral village, or the Kafkaesque lifeworld of modernity, or the commodified something-or-other of capitalism—is well-seen by a pastoral of a Kafkan or a capitalist way of seeing.[57]

This is what Ngai calls, in her writing on tone, the 'abstracting and generalizing' or 'virtualizing' nature of structures like mood, tone, and ideology: while a pastoral mood is in some sense a material structure particular to a meadowland, and a Kafkaesque tone is in some sense a material structure particular to the fictional world of 'The Trial,' and a commodity fetishist ideology is in some sense a material structure particular to a capitalist market, and in another sense an interpretive affective judgment of the relevant material structure, each of these structure is also a worldview directed at reality at large. To say it a bit loosely, a pastoral mood is how the structure of a meadowland calls on us to see reality at large, a Kafkaesque tone is how the structure of Kafka's imaginative artifacts calls on us to see reality at large, and commodity fetishism is how the structure of a capitalist market calls on us to see reality at large. A pastoral mood is an interpretive stance toward the cosmos, not just to a meadowland, just as a Kafkesque tone is an interpretive stance toward life, not just toward Josef K.'s fictional world or Prague, and a commodity-fetishist ideology is not restricted to the interpretation of a capitalist market. The inclusion of the 'commodity-fetishist ideology embodied in a capitalist market'—a cognitive schema for the interpretation of economic activity that models commodities as quasi-agents possessing motility—in this company, specifically, provides us with a good occasion to discuss an important facet of

[57]This may, I think, be part of what and Black studies scholar Fred Moten means when he writes that 'Blackness' as a poetics or a subject position has only a 'paraontological' relationship to the material, social, and psychical realities of Black individuals.

our broader understanding of the concept of 'ideology' as Ngai employs it. By 'commodity-fetishist ideology,' we here mean a schema for the interpretation of a cognitive schema for the interpretation of economic activity that models commodities as quasi-agents possessing motility. On at least some interpretations of Marxian philosophy, practical navigation in the material reality of a capitalist market does in fact require forming a commodity-fetishist representation of the market—it is not a pure form of misapprehension, but rather a schema which effectively maps the market while obscuring the contingent social conditions that make the capitalist market possible. I take Ngai's conception of ideology—'an imaginary relationship to a holistic complex of real conditions'—to operate largely in this narrowly Marxian milieu, where 'ideology' refers specifically to (cultural and psychological) cognitive adaptations to a social-material ecosystem that conceptually reify its contingent structure, rather than to any and all cognitive and affective phenomena downstream from 'power.' It is in this sense that, as Adorno writes about the ideology of bourgeois social science, 'false consciousness is also true':

> 'The separation of society and psyche is false consciousness; it perpetuates conceptually the split between the living subject and the objectivity that governs the subjects and yet derives from them. But the basis of this false consciousness cannot be removed by a mere methodological dictum. People are incapable of recognizing themselves in society and society in themselves because they are alienated from each other and the totality. Their reified social relations necessarily appear to them as an 'in itself.' What compartmentalized disciplines project on to reality merely reflects back what has taken place in reality. False consciousness is also true.' (1967)

There is a great deal more to say about this sense of 'virtual,' as we will understand it in this study, but for now let us consider Ngai's further characterization of structures like tone, mood, and ideology as 'immanent' and 'diffused' as well as 'virtual.' Not everything that's 'virtual' in a roughly Deleuzian sense should be comparable to 'virtual, diffused, and

immanent' structures like tone, mood, and ideology: some schema-systems we use to articulate reality, such as phonology or mathematics, deal in phenomena that aren't 'diffused' or 'immanent'[58] in the relevant sense. Tone, mood, and ideology, per Ngai, are '*immanent*' in that we cannot abstractly describe a given tone, mood, or ideology apart from from the material or sensate world that embodies it: it is a kind of 'heresy of paraphrase' applied to the meaning of worlds. They are, additionally, '*diffused*' in that there may well be no individual concrete elements within a world that independently exemplify the tone, mood, or ideology this world's totality embodies. One useful way to think of Ngai's description of tone, mood, and ideology as 'diffused, immanent' structures, I would propose, is that a world's tone, mood, or ideology is a systemic structures that nevertheless depends directly on the surface textures of immediate experience, putting the structural and the experiential (the system-level properties and surface-level properties) together without mediation from concrete, discrete, 'object-level' structures and relations. 'Diffused, immanent' structures are, in other words, what we might call 'ineffable' or 'ambient' structures: a world-systemic pattern that depends directly on the surface textures of immediate experience is *too diffused for showing* and *too immanent for telling*, closed off to everyday and analytic modes of thought in spite of sitting squarely within the confines of cultural and technoscientific materialism. A 'virtual' phenomenon of the 'diffused, immanent' persuasion, then, is a particularly ambient or ineffable virtual phenomenon. Putting it all together, we take Ngai's proposition that tones, moods, and ideologies are all examples of 'virtual, diffused, immanent' structures to mean roughly that a tone, a mood, or an ideology is an ambient or ineffable world-embodied subjectivity.

Going by the above criterion, I would propose, the company of 'virtual, diffused, imma-

[58]The above sense of 'immanence' is prima facie distinct from the material immanence of all virtual phenomena.

nent' structures turns out to accommodate a host of further critical and theoretical tropes that, like mood, tone, and ideology, negotiate with the idea of ambient or ineffable world-embodied subjectivity. We might consider, for example Jameson's concept of the 'cultural logic' of a literary epoch, the Situationist's concept of the 'psychogeography' of a city, Bourdieu's concept of the 'habitus' of a field, Rancire's concept of the 'sensorium' of a social-aesthetic regime, and Williams' concept of the 'structure of feeling' of a period, as well as most scholarly concepts of 'style,' and the lay concept of a 'vibe.' In truth, the genealogies of many of these concepts, diverse as they are, plausibly hark back to a certain early 20$^{\text{th}}$ century moment within German philosophical sociology, Phenomenology, and the philosophy of science that saw thinkers as violently different as Georg Simmel, Heidegger, and Rudolf Carnap[59] take up Wilhelm Dilthey's concepts of 'Weltanschauung' ('worldview')—'a totality in which on the foundation of a perception of the world or a 'world-picture,' questions about the meaning and sense of the world are decided' (Makkreel 351)—and 'Lebensstimmung,' the 'life-mood' or 'attitude to life' that structurally generates it. Historically as well as formally, the concept of a 'life-mood' is a junction where epistemology, Phenomenology, and social theory must negotiate the terms of their potential convergence, hierarchy, hostility, or complementarity, giving each discourse a claim to indispensability while also challenging each discourse's claim to autonomy. Consider, for example, the titular object of Frederic Jameson's canonical monograph 'Postmodernism, or The Cultural Logic of Late Capitalism.' In his iconic study, Jameson describes postmodernism as a kind of formal universal of the late capitalist era, comprising, among other things:

> 'A new depthlessness, which finds its prolongation both in contemporary 'theory' and in a whole new culture of the image or the simulacrum, [and] a consequent weakening of historicity, both in our relationship to public History and in

[59]See Gottfried on Carnap, and on the post-Dilltheyan discourse at large. See De La Fuente on Simmel. See Gosetti on Heidegger.

> the new forms of our private temporality, whose 'schizophrenic' structure (following Lacan) will determine new types of syntax or syntagmatic relationships in the more temporal arts; a whole new type of emotional ground tone—what I will call 'intensities'—which can best be grasped by a return to older theories of the sublime, [and] the deep constitutive relationships of all this to [computer network] technology, which is itself a figure for a whole new economic world system.' (7)

This late capitalist 'cultural logic' is diffused through every level of Jameson's world of late capitalism, from the spatial organization of a city to the temporal organization of a novel, from the causal order of economic systems to the normative order of philosophical systems, from the cultural structure of public History to the psychic structure of private memory. Acting on objects and on subjects, on abstracta and on concreta, on realities and on impressions, Jameson's 'cultural logic of late capitalism' is an aesthetic form, a way of thinking, and a social-material structure all at once. In his discussion, Jameson first discovers his 'cultural logic' in the logic of an eminent late capitalist art-style represented by the work of Andy Warhol:

> 'The first and most evident [difference between Van Gogh and Warhol] is the emergence of a new kind of flatness or depthlessness, a new kind of superficiality in the most literal sense, perhaps the supreme formal feature of all the Postmodernisms to which we will have occasion to return in a number of other contexts.' (9)

Using this diagnosis of a late capitalist art-style as a springboard, Jameson quickly starts extrapolating a late capitalist form of experience or subjectivity he calls, after Lacan, a 'schizophrenic' subjectivity:

> 'Magritte, unique among the surrealists, survived the sea change from the modern to its sequel, becoming in the process something of a postmodern emblem: the uncanny, Lacanian foreclusion, without expression. The ideal schizophrenic, indeed, is easy enough to please provided only an eternal present is thrust before

> the eyes, which gaze with equal fascination on an old shoe or the tenaciously growing organic mystery of the human toenail.' (10)

The logic of this 'schizophrenic' subjectivity, in turn, becomes for Jameson a kind of image of the social-material logic of everyday life under late capitalism itself:

> 'What characterizes the newer 'intensities' of the postmodern, which have also been characterized in terms of the 'bad trip' and of schizophrenic submersion, can just as well be formulated in terms of the messiness of a dispersed existence, existential messiness, the perpetual temporal distraction of post-sixties life.' (117)

Reflecting on the three aforementioned figurations of Jameson's 'cultural logic of late capitalism' side by side, we might observe that what one reasonably means by 'cultural logic' works out very differently depending on whether we're speaking of the logic of post-sixties life, the logic of a late capitalist 'schizophrenic' subjectivity, or the logic of a late capitalist art-style. When we discuss the logic of a social-material domain like 'post-sixties life,' we are typically asking for some kind of generative model: we're asking for a model of the forces animating the lifeworld of the late capitalist subject and the interactions of these forces in the structural production of the everyday world. The 'cultural logic' of post-sixties life, then, is a name for the dynamics of post-sixties social-material reality. A discussion of the logic of a subject, on the other hand, typically asks not for a causal model of the world but for a hermeneutic model of the subject's own modeling of the world—the 'logic' we are after is the logic of the subject's point of view, not the material dynamics of a world—rather than for a generative model of any material phenomenon. The 'cultural logic of late capitalism' that we can discover in a the late capitalist 'schizophrenic' subject is, accordingly, a kind of worldview or conceptual scheme typical of subjects within late capitalism. Finally, when we discuss the logic of an art-style or art-epoch, we place priority on explicating a criterion that can distinguish works of art that fit the style in question from works of art that do not.

While there may be a dozen dozen ways to talk about the logic of Warhol's artistic style, I'd argue that they all begin with the extraction of an aesthetic range—the range of possible Warholian aesthetic objects—implicitly defined by Warhol's practice. The 'cultural logic of late capitalism' that we can discover in an art-style like Warholian pop art, then, is at least initially a name for an aesthetic range or category.

For all the above difference of kind between the social-material logic of post-sixties life, the cognitive logic of the 'schizophrenic' subject, and the aesthetic logic of Warholian style, however, Jameson's discussion of key late capitalist phenomena such as 'the multinational, highrise, stagflated city' interweaves these three figurations of late capitalism's 'cultural logic' in a seamless network of metonymical and metaphorical continuities that appears to bind them into one holistic logic:

> '[Postmodernist depthlessness] can be experienced physically and 'literally' by anyone who, mounting what used to be Raymond Chandler's Bunker Hill from the great Chicano markets on Broadway and Fourth Street in downtown Los Angeles, suddenly confronts the great free-standing wall of Wells Fargo Court. (...) This great sheet of windows, with its gravity-defying two-dimensionality, momentarily transforms the solid ground on which we stand into the contents of a stereopticon, pasteboard shapes profiling themselves here and there around us. (...) If this new multinational downtown effectively abolished the older ruined city fabric which is violently replaced, cannot something similar be said about the way in which this strange new surface in its own peremptory way renders our older systems of perception of the city somehow archaic and aimless, without offering another in their place?' (14)

What, we might ask ourselves, is the actual object of experiential knowledge in the urban pilgrimage that Jameson describes above? Is it the new architectural aesthetic of Wells Fargo Court? Is it the geographic-economic order of the 'multinational downtown'? Is it the 'schizophrenic' abolition of our older systems of perception of the city'? When we imaginatively grasp Jameson's urban depthlessness, are we grasping a property of city life,

or a property of the late capitalist subject's mind, or a property of urban spectacle? Any specific answer to these questions should, I trust, strike us as arbitrary and reductive. The structure—an *ambient* structure, by our definition—that Jameson calls 'the cultural logic of late capitalism' is precisely what stays constant as Jameson freely alternates between material, cognitive, and aesthetic structures of late capitalism, as well as what enables Jameson to frequently leave the objects of his discourse hanging undetermined between the material, cognitive, and aesthetic realm with no loss of meaning. It is a kind of vanishing point, 'at the very edge of semantic availability,' (Williams 131) where a world's intertwined aesthetic, cognitive, and material structures come together as one ontologically multifarious regime.

It would be more or less a literary-critical truism, I believe, to say of writers like Melville or Flaubert, Stein or Musil, Platonov or Woolf, Beckett or Delaney, Pound or Acker, Ashbery or Goethe, that a portion of their work goes all-in on the prospect of expressing a 'virtual, diffused, immanent' structure in this sense—that is, the prospect of producing an aesthetically intelligible model of a psychological, social, cultural, phenomenological, or material world as a system. In literary works like 'Moby-Dick' or 'The Waves,' 'Faust II' or 'Dhalgren,' 'The Making of Americans' or 'Bouvard et Pécuchet,' the formal impositions of the drive to represent a 'virtual, diffused, immanent' structure take clear precedence over the formal impositions of narrative, rhetorical, and lyrical forms of structural coherence, such that we must conclude that the expression of the virtual, diffused, and immanent is not entirely within the purview of narrative, rhetorical, or lyrical strategies of representation.[60] This section will suggest that at least certain works of literature engaged in this kind of representation do so by constructing (or comprising) an imaginative or otherwise artefactual world whose idealized instantiation of a 'virtual, diffused, immanent' structure aesthetically

[60]Franco Moretti, paying particular attention to the scale of the reality in question, will cast many of these works as 'Modern Epics' seeking after the totality of the world system in the age of empires.

models an otherwise ineffable 'virtual, diffused, immanent' structure of the real world—that is, that at least certain works in literature engaged in this kind of representation do so through the cognitive-aesthetic construct we have called 'ambient meaning.' The object of a literary work's 'ambient meaning,' we suggest, can be variably understood as a structure of subjectivity (a *mood*), a logic of formal affinity (a *style* or *vibe*),[61] and a systemic model of social-material reality (a *system*), and—finally—as their interdependence in what Raymond Williams calls 'a structure of feeling.'

Our three-headed mathematical-aesthetical interpretation of ambient meaning—that is, sensible representation of an autoencoding manifold in an intersubjective input-space—as sensible cognition of a *mood*, as sensible cognition of a *style/vibe*, and as sensible cognition of a *system* makes intensive use of the sum total of conceptual machinery this dissertation introduced, and constitutes the highlight of this dissertation's project. Because of these considerations, we will now lay out the model in condensed, explicit form from an extreme bird's eye view, before we take a long walk through this same terrain from the frog's perspective. We begin by asserting three divergent theses on the object of ambient meaning:

A) *A (Modernist) work is a subjectivity.* The object of a literary work's ambient meaning (and thereby the major cognitive-aesthetic object of a 'constructivist' Modernist work) is a cognitive-affective structure close to what Heidegger calls a 'Stimmung.'[62]

B) *A (Modernist) work is an affinity of disparate materials.* The object of a literary work's ambient meaning (and thereby the major cognitive-aesthetic object of a 'constructivist' Modernist work) is a weak coherence or affinity in a field of cultural-material stuff.[63]

[61] We will make frequent use of the informal concept of a 'vibe' to highlight the inclusion of style-like structures or logics that are neither exclusive to form as opposed to content, nor exclusive to products of craft, in our concept of 'style.'

[62] Minimalist paraphrase of 'A': 'a (Modernist) work is a way of seeing.'

[63] Minimalist paraphrase of 'B': 'a (Modernist) work is an affinity of disparate materials.'

C) *A (Modernist) work is a cognitive mapping.* The object of a literary work's ambient meaning (and therefore the major cognitive-aesthetic object of a 'constructivist' Modernist work) is the systemic logic of a lifeworld.[64]

We broadly identify the object of '**A**' with the affect studies (Flatley, Ngai, Berlant) lingua franca concept of '*mood*,' which is in effect 'Stimmung' without the exegetical baggage. A cultural-Phenomenological approach to literary works, and to Modernist works in particular, as aesthetic 'anatomies' of culturally specific or politically significant moods is a thriving paradigm in affect studies, and arguably underlies the lion's share of robustly 'constructivist'[65] readings of Modernist works in contemporary literary studies. We identify the object of '**B**' with the loose, worldly kind of aesthetic unity everyday English calls a '*vibe*,' as well as with the object of a certain 'constructivist' approach to literary *style*, where form and content both collapse into a generalized principle of aggregation or selection:

> '[Ashbery's style] is both a comportment of Ashbery's, and a collection of images and ideas that assert themselves thematically across the course of his career' (Nealon 74)

> 'The achievement of many passages in The Waste Land is to make lines identical with lines Shakespeare or Webster wrote sound like lines Eliot might have written himself' (Kenner 79)

> '[Style] here is not a term for appropriateness of form to content, or for mere description of how language functions in a given text, but something *functional*, where the latter description not only identifies a replicable system (...) but also *generates* that replication.' (Hamilton)

[64] Minimalist paraphrase of 'C': 'a (Modernist) work is cognitive mapping.'

[65] Recall that 'constructivist,' in the sense that we have designated, requires not only a stance of 'radical artifice,' but also strongly constructive, rather than deconstructive, cognitive-aesthetic content. Ngai's reading of 'The Confidence Man' in 'Ugly Feelings' is, in our view, a model constructivist reading of a stereotypically (proto) Modernist work. Ngai's other readings in 'Ugly Feelings' have a slightly different orientation, focused on the dialectics of the reader's affective and ethical relationship to the literary work's mood.

Finally, we identify the object of '**C**' as a '*system*,' getting our money's worth from the ambiguity between the ontological and the epistemological meaning of 'system.' A sensible representation of the systemic logic of a lifeworld is, of course, identical with the 'aesthetic of cognitive mapping' of Jamesonian legend, putting this work at risk of literary-philosophical cryptozoology. Going for broke, we will model our discussion of this thesis on the Symbolist account of Modernist (Symbolist) texts as invocations of the 'correspondences' or 'esoteric affinities' that structure the perceptible surfaces of the world in accordance with 'primordial Ideals.'

On my account, each of the above three paradigms identically identifies the 'meaning' of a Modernist work with a trained autoencoder's manifold, describing one and the same formal structure via different functional definitions. The integrated story of our three 'Modernist paradigms,' under the manifold interpretation, can be told roughly as follows. To begin, let us recall that in unsupervised representation learning, we take a large set of data and learn a set of features that allows us to compressedly represent each piece of data in the set with minimal loss. We can regard this set of features as a 'mental language,' in a certain strictly heuristic sense: a trained net's low-dimensional representation is functionally comparable to sequences of qualified predicates—e.g. 'somewhat p, not f, not q, barely t, very g...'—such that the predicates stay constant from input to input but the qualifiers vary. To the extent that a deep dimensionality reduction is both necessary and effective for some given dataset, the 'vocabulary' of the language must approximately track the factors of variation entangled in the data. The features, in other words, must more or less correspond to something like Plato's objects whose ineffably interacting shadows on the cave's wall are the data—though this is a Platonism 'turned on its feet.' At the same time, the 'semantics' of the resulting 'vocabulary' is also likely to be highly non-modular, analytically intractable,

and 'situated.' With these considerations in mind, calling the features learned in deep dimensionality reduction '*esoteric affinities of perceptible surfaces to primordial Ideals*,' as in **C**, requires little if any poetic license.

Let us also recall, now, that the feature list extracted by unsupervised deep learning is equivalent to coordinates on a lower-dimensional manifold in the input-space, such that given the feature list can we efficiently ask of new data whether it lies close to this manifold, and with certain further allowances even generate random new data that lie on this manifold. While points on this manifold are not necessarily *similar* to one another, they collectively have an *affinity* to one another: as objects generated from the same highly constrained generative 'language,' they share at least some minimal form of what we might tentatively call a 'style' or 'vibe' or 'genre.' Equivalently, any sufficiently large—relative to the learning algorithm's complexity—set of data-points covered by the manifold has an unusually low absolute Kolmogorov complexity (i.e. has high compressibility), a condition which most sciences regard as necessary, but only ambiguously sufficient, for possessing a 'meaningful structure.' These are the promising, but in and of themselves '*weak*' fields of coherence that we know from '**B**.'

Finally, recall that once the training of an unsupervised deep learning algorithm is finished, when the algorithm encounters new input data it projects the data into the lower-dimensional manifold (within the input-space) corresponding to the feature list. The algorithm 'keeps' only the aspects of the data that are captured by the feature list, then treats these feature values as coordinates on the manifold, and then picks the corresponding item on the manifold as the reconstruction of the input data. I propose that quite apart from any implication that a trained unsupervised deep learning algorithm is a good overall analogue for the mind of a human subject, we have good (if provisional) grounds to hold that the

problem of locally linear dimensionality reduction is a staple of the 'ecological niche' that humans and learning algorithms for 'AI domains' share. If this is so, then we sometimes have reason to regard a person whose cognition is sensitive to some patterns in the world around her and insensitive to others—a person who excels at tracking certain similarities and differences but erases others—as projecting the worldly things she encounters into a lower-dimensional manifold spanned by her already-trained mental 'language.' On this account, we might treat the lower-dimensional manifold we're learning when our minds are doing a dimensionality reduction on the data in e.g. a Modernist novel as a 'reverse engineering' of the lower-dimensional manifold covered by the author's or narrator's 'mental language' (i.e. feature space). Or better yet, we might treat this manifold as the 'mental language' of a new way of seeing for which the novel's aesthetic meaning is a blueprint. We have arrived, in any case, at one form or another of a *Stimmung* as in '**A**.'

Let us sum up. I propose that Modernist works are often in the business of teaching you the coordinates of a lower-dimensional manifold in the space of possible data. When we look at the three paradigms for thinking about the cognitive product of Modernist works, '**C**' ('Ideals') treats learning the manifold as learning the generative model of the data, '**A**' ('Stimmung') treats learning the manifold as learning the dimensionality reduction method that produced the data as 'reconstructions' of other data, and '**B**' ('weak coherence') treats learning the manifold as learning that a certain set is compressible. In other words, the same low-dimensional ribbon in the input-space can be interpreted as the expression of someone's dimensionality reduction method (**A**), or as a set of objects that compress together well en masse (**B**), or as a set of objects such that the difference between any object y and any object x is a systemically meaningful difference (**C**).

And now, once more with feeling.

3.1 Mood

In 'Affective Mapping: Melancholia and the Politics of Modernism,' Jonathan Flatley places the representation and analysis of moods—both 'life-' regular, as we would have it—at the center of Modernism's cultural-political project. Flatley describes mood, after Heidegger's Stimmung, as a 'kind of atmosphere or weather' of cognitive and affective life whose rules of congruence delimit the space of possibilities within which 'intentions are formed, projects are pursued, and particular affects can attach to particular objects':

> 'Moods [Stimmungen] are the fundamental ways in which we find ourselves disposed in such and such a way. Moods are the how according to which one is in such and such a way' (FCM, 67). On the level of Stimmung, as Michel Haar writes, 'the world presents itself as what touches us, concerns us, affects us.' (...) Moods are not in us; we are in them; they go through us. ('It is not at all 'inside' in some interiority, only to appear in the flash of an eye; but for this reason, it is not at all outside either' [FCM, 66].) They 'assail us.' And in this sense mood is also total, or totalizing. Moods do not shed light on some one thing in particular, but on a whole environment: 'Stimmung imposes itself on everything' (66). Any orientation toward anything specific requires a presumed view of the total picture, a presumption that is usually invisible to us—that is just the way the world is.

While philosophically robust, Flatley's account leans more on literary works that dramatize the *consequences* of this Heidegerrian predicament—works we can read as parables about the social, psychopathological, and political consequences of the reign of an invisible presumption over our phenomenologies—than on literary works that labor to depict the niceties of a mood's logic, or the ontological and epistemic puzzles of mood. Of all the entities that make a human psyche, moods are perhaps the most intrinsically ethereal. States of mind like earnestness, nostalgia, melancholia, and estrangement are too abstract and 'syntactical' to belong to the order of emotions or perceptions, and too diffuse and open-ended to belong to the order of actions. In the first instance, a mood is determined not by having this or that

particular emotion or perception but by their dynamic and interplay over time: the kinship between the experience of melancholically eating breakfast and the experience of melancholically driving a car lies not in shared materia but in a commonality of structure, pattern, rhythm. In the second instance, mood has a free interplay with agency and action: a mood neither necessitates any particular action nor (constitutively) follows from any particular action.

The above 'paradoxical' nature of mood as a diffuse, distributed and syntactical pattern of subjectivity that is at once also the subject's structuring structure[66] and ur-experience is a recognizable thematic staple of Modernist ars poetica: one thinks of philosophically-minded authors ranging from Baudelaire and Proust to Stein and Ashbery who puzzle over the relationship between mood as a pattern of experience and mood as the experience of a pattern (cf. Benjamin's 'On Some Motifs in Baudelaire'), or of Conrad's unnameable epochs of sensation, or of Eliot's formula of poetry as 'concentration of a very great number of experiences which... would not appear to be experiences at all.' In considering mood as an object of representation, it's illuminating to consider the relationship of mood to the 'literal' constituents of subjectivity: emotions, perceptions, actions.[67] Mood, while primary from the Phenomenological or causal point of view, is *oblique*[68] to the language of actions and

[66]The phrase 'structuring structure' comes from Bourdieu. The phrase has no formal definition but intuitively refers to an abstract structure that has generative implications.

[67]One should here consider emotions, perceptions, and actions in the most expansive possible interpretation of each, encompassing intentions, active thoughts, subtle intimations, and whatever else.

[68]In calling moods 'oblique' I mean as follows: a) The mood facts about a person x globally and anomalously supervene on the totality of action, perception, and emotion facts about x. b) The mood facts about a person x depend on the fine details of x's actions, perceptions, and emotions, such that a list briefly naming the actions, perceptions, and emotions experienced by x (up to a limited level of specificity) may still allow for x's being in radically different moods. c) The mood facts about a person x depend on (abstract) similarity relations between different episodes of action, perception, and emotion (that possibly depend on their fine details), such that a detailed description of a single episode of action, perception, and emotion experienced by x may still allow for x's being in radically different moods.

sensation.[69] And while the language of action and sensation may have no special claim on the ontology of the human psyche, it has an obvious claim to being the default *intersubjective space of representation* for the human psyche—almost by definition, a 'factual' or 'uninterpreted' psychological description is a corpus of actions and sensations. States of mind (or ways of being) like earnestness, nostalgia, melancholia, and estrangement are too abstract and 'syntactical' to belong to the order of emotions or perceptions, and too diffuse and open-ended to belong to the order of actions. In the first respect, a mood is determined not by having this or that particular emotion or perception but by their dynamic and interplay over time: the kinship between the experience of melancholically eating breakfast and the experience of melancholically driving a car lies not in shared materia but in a commonality of aspect, structure, pattern, rhythm. In the second respect, mood has a free interplay with agency and action: a mood neither necessitates any particular action nor constitutively follows from any particular action. At the same time, because mood is here understood exactly as a structuring principle of experience and behavior,[70] the substance of a mood is just a corpus of perceptions, emotions, and actions.

Happily, as literary theorists we are by definition free to put aside some of the more metaphysical perplexities of mood in the ontology of mind[71] and focus on the structural perplexities of mood in the ontology of the *literary representation* of minds, exploring the representation of mood in textual media as a learning-theoretical[72] phenomenon. Indeed,

[69]Read 'sensations' in the most expansive sense, as shorthand for 'emotions, perceptions, and actions.'

[70]Since we are dealing with representations of moods rather than the things themselves, we can can abstract away the metaphysical question of whether mood has a phenomenology beyond that of its concrete instantiation.

[71]Chapter 1-2 of Charles Altieri's 2001 book 'The Particulars of Rapture' offer a good analytic inquiry into the engagement of art and poetics with the various metaphysical puzzles of mood qua metaphysical puzzles. The present discussion is largely congruent to Altieri's when overlapping. The introductory chapter of Thomas Pfau's 2008 'Romantic Moods: Paranoia, Trauma, and Melancholy, 1790–1840' offers an excellent historical treatment of the metaphysics of mood.

[72]By 'learning-theoretical phenomenon' we mean a phenomenon related to the theoretical questions that

we will happily takes our slogans for mood equally from Heiddegger and from the logical positivist Rudolf Carnap, arguing that in a literary text the two definitions—the phenomenological and the stylistic—are convergent:

> 'Moods are the 'presupposition' for, and 'medium' of, thinking and acting. (...) The mood has already disclosed, in every case, Being-in-the-world as a whole, and makes it possible first of all to direct oneself towards something.' (Heidegger, p.176)

> 'We find that metaphysics arises from the need to give expression to a man's attitude in life, his emotional and volitional reaction to the environment, to society, to the tasks to which he devotes himself, to the misfortunes that befall him. This attitude manifests itself, unconsciously as a rule, in everything a man does or says. It also impresses itself on his facial features, perhaps even on the character of his gait. Many people, now, feel a desire to create over and above these manifestations a special expression of their attitude, through which it might become visible in a more succinct and penetrating way. If they have artistic talent they are able to express themselves by producing a work of art.' (Carnap)

In the ontology of literary texts, where the 'presupposition for thinking and acting' becomes a text's presupposition for generating assertoric content, and the 'life attitude' of a person[73] becomes a text's tone, a 'mood' becomes a structure that is both a text's 'vibe' or aesthetic style and a text's distribution of the possible. Consider Georg Büchner's prose work 'Lenz,' famously cherished in Celan's 'Meridian' as an ur-phenomenon of an unnameable *something* crucial to Celan's own sense of life. The cognitive-affective force of 'Lenz,' as many critics have observed, lies in a sense of the transcendence of a certain sui generis attitude or feeling over the flux of experience and emotion, and over the boundaries between experience and reflection, concreta and abstracta, character and text. In concrete terms, 'Lenz' has a legacy for holding a consistency of mood across vivid episodes of sublimity and abjectness,

drive machine learning.

[73]'Person' here can come into play as author, narrator, focalization, character. We will later complicate the 'person' part of both these definitions by considering 'objective moods.'

quietude and turmoil, routine and catastrophe, and across different ontological levels of discourse in the narratological sense. As critic David Auerbach writes, 'Lenz moves over the course of the story from revelation to agony and shutdown. It's not clear that they are any different for him; his revelations have the same visceral force as the pain.' So do most differences between mountains and rooms, conversations and dreams, friends and objects, ideas and sensations, narrator and character, minutes and days give out, as the Lenzian gestalt imprints itself on the raw stuff on the world:

> 'Everything seemed so small, so near, so wet, he would have liked to set the earth down behind an oven, he could not grasp why it took so much time to clamber down a slope (...) Only sometimes when the storms tossed the clouds into the valleys and they floated upwards through the woods and voices awakened on the rocks, like far-echoing thunder at first and then approaching in strong gusts, sounding as if they wanted to chant the praises of the earth in their wild rejoicing.'
>
> ('Es war ihm alles so klein, so nahe, so naß; er hätte die Erde hinter den Ofen setzen mögen. Er begriff nicht, daß er so viel Zeit brauchte, um einen Abhang hinunter zu klimmen, einen fernen Punkt zu erreichen; er meinte, er müsse alles mit ein paar Schritten ausmessen können. Nur manchmal, wenn der Sturm das Gewölk in die Täler warf und es den Wald herauf dampfte, und die Stimmen an den Felsen wach wurden, bald wie fern verhallende Donner und dann gewaltig heranbrausten, in Tönen, als wollten sie in ihrem wilden Jubel die Erde besingen...')
>
> 'They continued talking, he searched for words and they came tumbling out, but it was torture; little by little he calmed down, the cozy room and the tranquil faces looming out of the shadows, the bright face of a child on which all the light seemed to rest, trusting eyes raised in curiosity, and finally the mother sitting quietly back in the shadows, angel-like.'
>
> ('Man sprach weiter, er suchte nach Worten und erzählte rasch, aber auf der Folter; nach und nach wurde er ruhig—das heimliche Zimmer und die stillen Gesichter, die aus dem Schatten hervortraten: das helle Kindergesicht, auf dem alles Licht zu ruhen schien und das neugierig, vertraulich aufschaute, bis zur Mutter, die hinten im Schatten engelgleich stille saß.')
>
> 'Yet the more he grew accustomed to his life, the calmer he became, he helped out Oberlin, sketched, read the Bible; old vanished hopes rose anew in him; the

> New Testament spoke so directly here, and one morning he ventured forth. When Oberlin recounted how an invisible hand had steadied him on the bridge, how his eyes had been dazzled by a blinding light on a mountain, how he had heard a voice, how it had spoken to him in the night...'
>
> ('Doch je mehr er sich in das Leben hineinlebte, ward er ruhiger. Er unterstützte Oberlin, zeichnete, las die Bibel; alte, vergangne Hoffnungen gingen in ihm auf; das Neue Testament trat ihm hier so entgegen... Wie Oberlin ihm erzählte, wie ihn eine unsichtbare Hand auf der Brücke gehalten hätte, wie auf der Höhe ein Glanz seine Augen geblendet hatte, wie er eine Stimme gehört hätte, wie es in der Nacht mit ihm gesprochen...')

The form of Lenz's speech is close to the form of his thoughts, which is close to the form of his days at the parsonage, which is close to the form of his paces on the mountain, which is close to the form of his gaze, which is close to the form of his pains and ecstasies, which is like the form of his attention in a minute's span. The Jakobsonian *dominant* of Lenz is, in this sense, marked both by the pull of the distributed, global pattern of Lenz's constancy and by the salience of the structural form of each Lenzian lyrical moment. In one sense, what is crucial in the transformation of Lenz's lyrical processes into a *mood* is that the Lenzian affect subsist in the *global* relationship of the episodes' structural forms to one another, rather than in each particular lyrical episode. But in a deeper sense it is spurious to even discuss the local structural forms as something separate from their global affinity: it is exactly from the global view that we discern an abstract affinity or invariance across episodes, and comprehend each episode as instantiating an invariant structural form in variant materia. The mood-making in 'Lenz,' in this sense, strikingly exemplifies how in mood the structural dimension emerges from the immanent in the construction of a mood—especially because it is iteration of Lenz's hyper-bodily form of perception that becomes a timeless, abstract mood-pattern. As Edmunds[74] usefully observes, 'Lenz' strips Lenz's agency of narrativity

[74]Edmunds 180.

by iterating every dramatically pregnant action, whether external or internal, twice or more. In parallel, it strips the immediate experiential world of Lenz's perceptions and emotions of narrativity through a ubiquity of cyclical rhythms and phenomena, both natural and cultural, physical and mental. The tendency to the condition of combinatorial exhaustiveness in 'Lenz,' wherein sequences of action are repeated at a permutation, and every perception has an equal and opposite perception, shifts the ground from the psyche as a process to the psyche as a combinatorial space or a vocabulary. It would be crude but plausible, I think, to argue that when we read fiction we construct psychological models partly or even primarily by asking why, out of several relevant alternatives $a, b, c...$ the protagonist ended up doing (or saying or believing) a and not b or c. The exhaustiveness of 'Lenz' largely collapses the distinction between what Lenz can meaningfully think and do and what Lenz thinks and does. If a mood is, as Heidegger had it, 'the 'presupposition' for, and 'medium' of thinking and acting,'[75] 'Lenz' replaces parsing thoughts and actions with parsing the contours of the space of thoughts and actions accessible to a given medium of thinking and acting.

The aesthetic form of 'Lenz' as a *constrained space of possibilities* and the *limits of the space of thoughts and actions* accessible to Lenz's 'medium of thinking and acting' are, perhaps, easily seen to be two ways to conceptualize the same phenomenon, to the extent that they both succed in conceptualizing at all. From the first point of view invariance is understood in a 'top-down' way as a regularity structuring the text, and from the second point of view invariance is understood 'bottom-up,' as emergent from the organic dynamics of a mood that generates the text.[76] Moreover, we don't need to look particularly hard

[75] Heidegger 90.

[76] Above, as elsewhere in the present chapter, our discussion of a 'generative' structure leaves aside distinctions between the author literally generating the text and the protagonist figuratively generating the text. We leaves this distinction aside because, plausibly, the author's generative model of the protagonist's mood generates the text in both senses.

to find use for the mathematical trope that would conceptualize of this mood-space as a trained autoencoder's lossy compression schema, where so much of Lenz's waking-dream-work is work of *reduction and selection*, luminously apprehending or disclosing structures and relations of a certain ontological type while matting or abstracting away everything orthogonal to these structures.[77] And the syntactic anomalousness of 'Lenz' discloses itself on a sentence-by-sentence basis; it is only on the long, *ambient* view of the text, enabled by the foregrounding of repetition, verbal echoes, parallelism and antitheses, that the radical *semantic* landscape of 'Lenz'—and of Lenz's mood—discloses itself. The madness is largely in the method. Consider the following randomly chosen selection of sentences from 'Lenz':

1. 'He felt no sense of fatigue, except that sometimes it annoyed him that he could not walk on his head.' ('Müdigkeit spürte er keine, nur war es ihm manchmal unangenehm, daßer nicht auf dem Kopf gehn konnte.') 2. 'He'd have liked to set the earth behind the stove to dry.' ('Er hätte die Erde hinter den Ofen setzen mögen.') 3. 'The full moon hung in the Heavens; the locks of his hair fell across his temples and his face, tears hung on his eyelashes [...] and the moon shone the whole night and hung there above the mountains.' ('Der Vollmond stand am Himmel; die Locken fielen ihm über die Schläfe und das Gesicht, die Tränen hingen ihm an den Wimpern und trockneten auf den Wangen—so lag er nun da allein, und alles war ruhig und still und kalt, und der Mond schien die ganze Nacht und stand über den Bergen.') 4. 'The world he had once hoped to serve was severed by a monstrous gap, he had [...] only a dreadful void in himself.' ('...die Welt, die er hatte nutzen wollen, hatte einen ungeheuern Riß; er hatte... eine

[77]Lenz's radical form of apperception is not principally that of a Bretonian oracle sublating the observable unto a sur-reality, or that of a Coleridgeian transformative imagination, despite occasional forays into both. Recall Lenz's invective against idealism in the 'Kunstgespräch' section—Lenz is by nature no friend to the dominion of mind over reality. As Horton argues in his literary-linguistic study of 'Lenz,' there is in fact a relative paucity of strong tropes in the text. Müller-Sievers similarly identifies Lenz as a non-romantic or anti-romantic text on this account, rejecting the interpretation of the text's iconic prima facie semantic deviation ('... except sometimes it annoyed him that he could not walk on his head') as an adynaton or any other rhetorical non-literality.

schreckliche Leere...') 5. 'At times great masses of light rose from the valley like a flood of gold.' ('Gewaltige Lichtmassen, die manchmal aus den Tälern, wie ein goldner Strom, schwollen, dann wieder Gewölk, das an dem höchsten Gipfel lag und dann langsam den Wald herab in das Tal klomm...') 6. 'Everything was empty and hollow to him.' ('...es war ihm alles leer und hohl...') 7. 'Clouds swept rapidly across the moon.' ('Wolken zogen rasch über den Mond...') 8. 'This mountain of madness shot up at his feet...' ('Jetzt wuchs sie, der Alp des Wahnsinns setzte sich zu seinen Füßen...') 9. 'The landscape [...] was so narrow he was afraid he would bump into everything.' ('...die Landschaft beängstigte ihn, sie war so eng, daßer an alles zu stoßen fürchtete.') 10. 'It seemed to him that he lay in an infinite sea that gently rose and fell' ('...es war ihm, als läge er an einem unendlichen Meer, das leise auf und ab wogte.') 11. 'The world was an open wound to him.' ('Das All war für ihn in Wunden...') 12. 'He amused himself by standing houses on their roofs.' ('Er amüsierte sich, die Häuser auf die Dächer zu stellen...')

In Lenz's world qualities and predicaments of every sort are measured largely in terms of moving things around (1, 2, 12), things moving him around (8, 10), things moving around (5, 7), orientation (1, 3, 5, 10, 12), tissue damage (4, 11), and pressure (3, 4, 6, 9), making no real distinction between his body and the external world on each count. I want to understand this prevalence of spatial- corporeal language permeating 'Lenz' as an emergent characteristic of the feature-space of Lenz's apperception. If we think of Lenz's apperceptive space as a lower-dimensional manifold in perception- space, where each small perceptual difference that makes an apperceptive difference corresponds to a movement tangent to the manifold and each small perceptual difference that does not make an apperceptive difference corresponds to a movement orthogonal to the manifold, we can ask to which of the common dimensions of perceptual variation are the manifold's coordinates reliably sensitive or insensitive, to various degrees.[78] Lenz's apperception is acutely sensitive to interoceptive

[78]If the answers to such questions were too systematic, rather than rough, relative, and heuristic, we would

variation (the senses of internal bodily states), to proprioceptive variation (the senses of the relative position of one's body parts), to chronoceptive variation (the sense of duration) and to vestibular variation (the sense of acceleration and direction), and less sensitive to extroceptive variation (the five standard senses), with the exception of degrees of brightness and of loudness. The lost extroceptive richness in Lenz's apperception is compensated for by an acute apperceptive sensitivity to the play of the many 'echoes' that one's encounters with an external world continuously cast unto one's inwardly-directed sense- modalities: as Phenomenologists and phenomenologically-inclined empirical psychologists often stress,[79] perception is replete with affordances (simulated corporeal interactions with objects) and with proprioceptive and vestibular simulations of the position and movement of objects, and Lenz apperceives these perceptual movements extremely finely with consistency. Yet even though these world-registering aspects of the self-directed modalities of perception are common within normal apperception, the restriction of Lenz's capacity to register the external world to these self-directed modalities collapses any real apperceptive distinction between external objects and Lenz's corporeal reaction to these objects. For Lenz the world is cracked and/or Lenz's body is hollow; Lenz wants to walk on his head and/or flip houses on their head; Lenz needs to warm up his cold and damp landscape by a stove; Lenz feels the weight of the tears hanging from his eyes and/or the weight of the moon hanging from the sky; Lenz is wounded by the world and/or has a wound in the world part of his body. Taking the Emersonian transparent eye to its logical conclusion, the external world is like a part of Lenz's body that can feel but cannot move at his will—an uncooperative set of limbs, or rather Lenz is like the limb of an uncooperative body. Body and world made one by

no longer be dealing with a mood (or any high-level abstraction at all.) I discuss the issue further below.

[79]The topic is very remote from the subject of this dissertation and should hardly be controversial in the present context, but see for example Kelly's 'Merleau-Ponty and the Body.'

sharing in a single space of representation, as if in an apperceptive synesthesia in which the 'modality' of the apperception of the self is extended over the domain of world-directed apperception, or the apperception of the world is projected into the modality of self-directed apperception.[80]

Crucially, the above does not go towards *taking stock* of Lenz's perceptual 'vocabulary'—I am not suggesting that the formula for Lenz's mood is to remove certain modalities—but rather goes towards *describing some emergent properties* of the manifold spanned by this 'vocabulary.' Or, in more technically sophisticated terms, we are describing some of the geometrical relationships of movements in Lenz's apperception-space to movements in the intersubjective apperception-space within which Lenz's apperception-space constitutes a submanifold. This apperception-space structuring 'Lenz' is in one sense a psychical property of its protagonist, but in another sense an aesthetic property of the text structured under Lenz's sign, constituting a sui generis aesthetic dimension of the work. 'Aesthetic,' but not 'formalist'—or, at least, not 'formalist' in any sense we could seek to contrast with content. The aesthetic dimension which extends from psychical mood in a work like 'Lenz' *partly* comprises linguistic constancy of pattern of the type that British literary linguists call 'stylistics,'[81] but subsumes linguistic constancy of pattern as raw matter into larger, omnivorous patterns that comprise occurrences, behaviors, objects, and environments. The literary representation of psychical mood thus verges on what one might informally call a text's 'vibe,' alluding as it does to something diffuse, global and abstract that gives the text its sensate identity.

[80]These structural phenomena complement and magnify, but are distinct from, the more classical Sturm und Drang metaphoric and symbolic correlation of natural phenomena and emotional states in the text. Indeed, the concretization of the trope in Lenz's apperception is congruent with Edmund's reading of Lenz as in part a parody of Goethe's Werther.

[81]See for example the work of Michael Halliday.

3.2 Style/Vibe

The mathematical trope of a trained autoencoder's canon, we propose, is key to understanding structures of meaning of the 'virtual, diffused, immanent' kind in a literary work, and arguably key to understanding the epistemology of 'virtual, diffused, immanent' structures at large. The major onus on this section is to make the case that a good share of what we would intuitively recognize as a literary work's aesthetic-cognitive dealings in 'virtual, diffused, immanent' structures (as defined by reference to paradigm cases like 'Godot' or 'Faust', 'The Trial' or 'The Waves', 'Bouvard et Pécuchet' or 'The Confidence Man') proceeds via the demarcation of an aesthetically structured set. One of the benefits of the autoencoding-infused theory of literary meaning that this chapter proposes, I hope, is the relative originality of a literary-critical perspective whose primary hermeneutic object is 'a set of objects or phenomena.' This very benefit, however, can also threaten to undo the 'literary' side of our promised literary-technical analogy between a literary work's imaginative set of objects or phenomena and an autoencoder's canon, given that the idea of 'sets of objects or phenomena' as primary aesthetic or ideological structures of meaning in a work of art is not an already-established literary-theoretic trope.[82] Nevertheless, I would propose that on reflection we have independently compelling reasons to conceive of the contents of some (and possibly most) literary works as each comprising, in part, an imaginative set of objects or phenomena. The leading intuition is, in short, that one major structure of meaning in a literary work is the aesthetic unity of a work's imaginative range—that is, the aesthetic unity of the set of all objects or phenomena spanned by the fictional or notional world that the work imaginatively constructs. The meaning of a literary work like Dante's 'Inferno,'

[82]Deleuzian assemblage are almost certainly closely related to this dissertation's topic, but a Deleuzian assemblage is a set with an explicit, rather than an implicit, relational structure.

Beckett's 'Waiting for Godot,' or Kafka's 'The Trial,' we would likes to say, lies at least partly in an aesthetic 'vibe' or a 'style' that we can sense when we consider all the myriad objects and phenomena that make up the imaginative landscape of the work as a kind of curated set. The meaning of Dante's 'Inferno,' for example, is in part in that certain *je ne sais quoi* that makes every soul, demon, and machine in Dante's vision of hell a good fit for Dante's vision of hell. Similarly, the meaning of Beckett's 'Waiting for Godot' is partly in what limits our space of thinkable things for Vladimir and Estragon to say and do to a small set of possibilities the play nearly exhausts. The meaning of Kafka's 'The Trial,' in turn, is partly in the fractal repetition of an abstract pattern we now use to ground the English predicate 'Kafkaesque.' These three ready-to-hand examples are, of course, a biased choice of literary works, serving to help arouse an intuition, but the underlying principle is plausibly broad in application (if not always equally evocative: part of the meaning of Bront's 'Villette' lies in the set of 'Villette'-conforming objects and phenomena; part of the meaning of Goethe's 'Faust' lies in the set of 'Faust'-conforming objects and phenomena; part of the meaning of Stein's 'Tender Buttons' lies in the set of (possibly inherently linguistic) 'Tender Buttons'-conforming objects and phenomena).

While, clearly, we have ways to go before proposing any structured correspondence between these still-suspicious candidate aesthetic structures ('the set of 'Faust'-conforming objects or phenomena,' 'the set of 'Godot'-conforming objects or phenomena,' and so on) and the mathematical trope of autoencoding, we might nevertheless consult our knowledge of autoencoders here to illustrate a broad point about the potential meaning of such 'sets of objects or phenomena.' Autoencoder algorithms, as we know, will learn the logic of a worldly domain by extrapolating from a training set of objects or phenomena sampled from the domain, and thereafter express this logic in the form of a specially curated 'canon' of objects

or phenomena. Autoencoders, we know, deal entirely in worlds rendered as sets of objects or phenomena, and so whatever deeper worldly structures come into play in autoencoding can evidently find expression in the structure of a set of objects or phenomena. Though neither works of art nor real-life worlds can, of course, ever wholly be described as sets of objects or phenomena, we argue that the set of all the objects or phenomena found in a given world, or found in the imaginative landscape associated with a given work of art, is a powerful proxy for a space of possibilities that underlies the various internal movements of a world or of a work of art. Speaking concretely, I'm proposing that by sampling objects or phenomena from a world, or from the imaginative landscape associated with a work of art, a subject or a reader extracts a collection of objects or phenomena that she may treat as an example of the *kind* of objects or phenomena the world or work in question can allow. The concept of a 'vibe' or 'style,' as we will understand it in this section, refers to what we grasp when we *aesthetically* grasp a collection of objects or phenomena as an example of the logic of a world or work. If the collection of objects or phenomena associated with a given world or with a given work possesses an 'aesthetic unity' of the right kind, we argue, then the world or the work in question has a humanly accessible (though likely analytically intractable) systemic grammar that determines what objects or phenomena are *possible* within the world or work, as well as the respects of similarity and difference that describe the variation of objects or phenomena within the world or work. Furthermore, we argue that this 'aesthetic unity' in the collection of objects or phenomena associated with a world or with a work acts as a schema for the systemic grammar of the world or work, allowing the subject or reader to grasp the space of possibilities that gives the objects or phenomena within the world or work their meaning.

In the course of the present section, we will work both to study this relationship between a

world or work's 'vibe' or 'style' and its systemic grammar for its own sake, and to leverage this relationship into the main argument of this chapter—the argument that the vibe or style of a literary work, as expressed in a set of imaginative concreta, can act as an aesthetic schema not just for the structure of the literary work's own imaginative world but also for the structure of a *real-life* world, and thereby constitute a kind of Symbol or embodied cognitive mapping. For now, however, we will hold off on the promised 'vibe/style as aesthetic Symbol' thesis, and keep our interest focused on the matter of a vibe's role in a literary work's *internal* sense and sensibility a little longer. While the idea of a literary work's 'horizontal' logic that we sketch above does not exactly lack for more established, literary-theoretically sophisticated cousins—Sianne Ngai, in particular, has studied many concepts of a literary work's 'global, virtual, diffuse, and immanent' affective structure, and the Deleuzian idea of an 'assemblage' is in part about the power of 'a set of myriad objects of phenomena' to create meaning—I want to hold on to a certain specificity that comes from thinking of this logic as a 'style'-like ('style' or 'vibe') aesthetic structure constituted by whatever 'objects or phenomena' make up a given literary work. Where the traditional literary-critical concept of 'style' primarily concerns the verbal logic of a literary text, and may even suggest concern with a work's presentational structure over the structure of whatever the work's presentational structure presents (when such distinctions are applicable), the objects or phenomena that collectively constitute a literary work's 'vibe' or 'style' in the sense above need not be verbal, and may well be constituents not of the text per se but of a world, a field of imagery, or network of ideas born of the text—all forms of what we'd call a literary work's 'imaginative landscape.' We have suggested above, for example, to conceive of Beckett's 'Waiting for Godot' as partially a set of actions that collectively define a *style* of action, and to conceive of Dante's 'Inferno' as, partially, a set of beings—the inferno's various souls, machines, and

demons—that collectively define a *style* of beings.

Though in some ways any generalized concept of 'style' as above owes more to many extra-literary aesthetic meanings of 'style'—the style that ties together an ensemble or a furnished living room, a musical style, a style of thought—than to the literary-critical concept of style, I nevertheless want to suggest that our idea of a 'vibe' or 'style' in the imaginative landscape of a literary work represents an interpretation and extension of the more traditional idea of literary style, rather than a mere homonym. Dealing with very different concerns from those that drive the present dissertation, Diana Hamilton's recent study of style, 'Style in Quotation Marks,' arrives at an idea of literary style that closely corresponds to what this dissertation would describe as a 'vibe' or 'style' of a set of sentences. Hoping to liberate the literary concept of 'style'—'the strange consistency that allows a text to take on the impossible specificity of the *author-turned adjective*'—from the romantic or modernist shackles that reduce it to an author's voice, Hamilton describes a literary text's 'style' as the minimal consistency that turns a text into an aesthetic system. For Hamilton, the question of 'style' concerns not the relationship between an author or a corpus and a work, but rather the relationship between the (material or imaginative) particulars that make up a literary work and the replicable pattern they express in rising to the level of an aesthetic system. A style, Hamilton writes, 'is, in its loosest meaning, the expression of the interaction between elements in a way that neither reduces the work to [its individual elements] nor loses sight of them.' On Hamilton's view, as on our own, the structure that expresses such an interaction is necessarily 'something *functional*, where the latter description not only identifies a replicable system (...) but also generates that replication.' What Hamilton proposes, I believe, is that we may conceive of 'style' as a structural *cause* of a work's elements, rather than only as an emergent aesthetic form. While this is by no means a precise technical distinction, the

broad idea of treating 'style' as a (replicable but ineffable) generative function expressing the aesthetic relationship of the elements it generates makes for a powerful conceptual intermediary between the ordinary literary-critical idea of 'style' and our concept of a 'style'-like structure as the aesthetic schema of a constrained space of possibilities.

If Hamilton extracts from the traditional literary-critical concept of 'style' the idea of a minimal consistency that makes a set of sentences legible as as an aesthetic system, then our concept of a 'vibe' or 'style' extends a like idea of consistency to the realm of 'objects or phenomena' at large, in order to apply it at whatever ontological registers best capture the horizontal self-organization of a given work. As we will see in the next chapter, wherein we devote a more detailed discussion to specific processes by which a literary text constructs a 'vibe'-defining set of objects or phenomena as its imaginative landscape, different literary works will construct 'vibe'-defining sets of objects or phenomena in different, and often multiple, ontological registers. The ontological register in which we will construe a given work of literature's imaginative landscape of objects or phenomena, and wherein we will discover a 'vibe' or 'style,' is, as a rule, an ad hoc adaptation to each literary work's form of self-organization: the best way to conceive of the imaginative landscape of a given literary work may be as a set of sentences, or a set of fictional events, or a set of plot-lines, or a set of images, or a set of affective states, or none or all of the above. A reader or a theorist can fruitfully conceive of the imaginative landscape of Musil's 'The Man Without Qualities,' for example, as a set of human types comprising the various characters introduced within the novel, or as a set of philosophical encounters comprising Ulrich's various negotiations with his interlocutors' lived philosophies, or as a set of cognitive-affective states which Musil's paragraphs depict, or as a set of German-language paragraphs. It could prove difficult, on the other hand, to successfully conceive of Stein's 'Tender Buttons' as a set of rooms, objects,

and foods rather than as a set of textual compositions, or to productively conceive of the imaginative landscape of a Bashō haiku as a set of verse lines rather than as a set of images or concepts connoted by the text. What introducing the new critical trope of a literary work's associated 'set of objects or phenomena' gives us, then, is the flexibility required to deal with the 'horizontal' structure of each literary work on its own terms, whether these terms are best configured in a textual-material, conceptual, narrative, imaginary, or affective register.

Having now hopefully established some beginnings of intuitive familiarity for the idea of 'virtual, diffused, immanent' aesthetic/ideological structures of meaning in a literary work, we can review this section's argument for an autoencoding-inspired model of these structures, and of their possible role in a 'cognitive mapping' or Symbol-like representation of real-life worlds. The present section argues that what makes the canon of a trained autoencoder a concrete expression of a method of mimesis and its matching worldview, or a mood or a gestalt, is a kind of 'aesthetic unity' between the objects of the canon in their intersubjective input-space form—an aesthetic unity closely related to what we intuitively call a 'style' or 'vibe.' Whatever point of view, gestalt, theory, mood, interpretation, or cognitive map the computational intricacies of a given trained autoencoder's feature function stand for, this same system of thought or abstract model necessarily has a corresponding formulation as a set of objects that collectively express a 'style.' This mathematical relationship between the 'style'-like aesthetic logic of a trained autoencoder's canon and the trained autoencoder's system of abstractions, we thus argue, is an important aspect of a literary work's capacity to ambiently ground complex systems of thought, or ambiently model complex worldly domains.

Formally speaking, our proposal in this section will consist of two distinct theoretical parts that each apply the framework of autoencoding to a question of aesthetics and rep-

resentation. In the first part of our argument, we propose an autoencoding-inspired model of the relationship between the aesthetic unity (the 'style' or 'vibe') of a set of imaginative objects or phenomena associated with a work of art and the systemic structure of the work of art's imaginative landscape. In the second part, we propose an autoencoding-inspired model of the capacity of the imaginative landscape of a work of art to aesthetically represent the systemic logic underlying a real-life world. Crudely, we might say that the first part of our argument concerns the question: 'what is it we sense when we sense that (e.g.) 'The Trial' has a vibe?' The second part of our argument, in turn, concerns the question: 'what is it we sense when we sense that this vibe expresses something about real life?' In each case, we propose that the conceptual framework of autoencoding respects and extends our best literary-theoretic understanding of aesthetic meaning-making. Furthermore, we will argue that our autoencoding model of a vibe's internal structure (our answer to the question 'what do we sense when we sense a vibe?') and our autoencoding model of a vibe's representational capacity (our answer to the question 'what do we sense when we sense this vibe expresses something about real life?') converge into a single cognitive-aesthetic hypothesis, making both models more attractive on grounds of explanatory parsimony.

On the hypothesis that will inform this section's answers to our two 'vibe' questions, the imaginative landscape of a literary work is functionally analogous to a sample from the canon of a trained autoencoder. In fact, I argue that once we reduce the mathematical trope of a trained autoencoder's canon to its most general form—a set whose manifold geometry is losslessly compressible—it may become superfluous to qualify its structural relation to a literary work's imaginative landscape as an analogy. A literary work's 'vibe,' per this theory, is a losslessly compressible manifold structure of the set of objects or phenomena that makes up a work's artefactual or imaginative landscape in a relevant intersubjective input-space,

and equivalently an approximate compression schema for worldly domains approximated by the manifold. The structure of this chapter's argument is, in this sense, a typical pragmatist 'inference to the best explanation': we argue that both our best informal insight with regard to the *internal structure* of a literary work's 'vibe' and our best informal insight with regard to the *representational* capacity of a literary work's 'vibe' are well-explained if a hypothesis h is true, therefore h. Arguments of this form are, of course, highly defeasible, but they are nevertheless a powerful—and possibly indispensable—means for assigning credence to hypotheses, especially in areas of inquiry that do not have a standard operationalization.

Over the course of the previous section, we began to think about the fit of a trained autoencoder's canon to a compact generative vocabulary as an intrinsic structural property of the canon, given directly by the canon's manifold geometry. Speaking in crude terms, we have seen that if a set of objects corresponds to a geometrically compressible manifold in an intersubjective input-space, this set of objects aligns with the constrained generative 'vocabulary' of some hypothetical trained autoencoder. More technically, we learned that we might therefore say this set of objects has a practicable *lossless compression schema*. I argue that we might think of this property, a little loosely at first, as the 'aesthetic unity' of all the objects that make up the manifold—a regularity of forms that makes it possible to subsume the variety of the manifold's objects into the interplay of a relative handful of structural parameters. This 'aesthetic unity,' I would propose, is closely related to the aesthetic unity we grasp whenever we are grasping the *style* or *vibe* of a collection of objects or phenomena, whether it is the style associated with the imaginative landscape of some literary work, or the style associated with the contents of some worldly domain. In other words, the act of grasping a 'style' or a 'vibe' in a collection of phenomena may be comparable to successfully training an autoencoder. Importantly, I'm not suggesting that all, or even many, worldly

collections of phenomena marked by consistency of 'style' or 'vibe' are *themselves* a kind of trained autoencoder's manifold. I am suggesting, instead, that the act of grasping a 'style' or 'vibe' in a collection of phenomena may be comparable to the construction of a geometrically compressible manifold whose objects approximate the phenomena in the collection—and, perhaps, that we might think about the mathematical totality of the resulting manifold as capturing this 'style' or 'vibe' in its Ideal or 'primordial' form, or at least Idealized form. Nevertheless, I argue, human life is likely not without collections the are, themselves, like a trained autoencoder, since where they don't exist we have good reason to create them.

If the comparison between autoencoding and grasping an aesthetic unity of 'style' or 'vibe' in a real-world collection of phenomena holds sway, then it might sometimes be right to think about a *work of art*, in turn, as the material realization of a given style in its Ideal or idealized form—or, more prosaically, as analogous to (a sample from) the image of the projection function of a trained autoencoder that models the real-world data marked by a unity of style well. On the account that I propose, a work of art can represent the 'aesthetic unity' of a natural collection of phenomena by presenting an artefactual collection of phenomena such that:

1. For every phenomenon x in the natural collection, some phenomenon y in the artefactual collection is an adequate approximation of x.

2. The set of all phenomena composing the artefactual collection losslessly aligns with a constrained generative vocabulary.

Suppose that when a subject grasps an 'aesthetic unity' in a collection of worldly phenomena—perhaps the collection of sights and sounds associated with a certain city, or the collection of behaviors associated with a certain social institution, or the collection of experiences associated with a certain life-predicament, or the collection of cultural-material artifacts associated

with a certain culture—part of what she grasps can be compared to an autoencoder trained on this collection. In other words, the essence of what the subject learns is comparable to the manifold-structure of a trained autoencoder capable of approximately reconstructing the phenomena of the collection. An autoencoder's manifold is ultimately itself just a (losslessly compressible) collection of phenomena—what we have called a trained autoencoder's '*canon*'—and so the 'aesthetic unity' the subject grasps in the *original collection of phenomena* has a direct formal expression as *another collection of phenomena*. Intuitively, the phenomena of the second collection are a kind of reinterpretation of the phenomena of the first collection in accordance with the first collection's own 'style' or 'vibe.' The phenomena of the autoencoding manifold are thus, quite literally, idealizations of the phenomena in the original collection: the autoencoding manifold comprises phenomena of the original collection reformed in accordance with an ideal derived from the original collection, which the original collection imperfectly embodies. We argue that, similarly, when the collection of objects or phenomena associated with a work of art possesses a strong unity of style, the aesthetic unity of the artefactual collection is potentially an idealization of a looser, weaker aesthetic unity between the objects or phenomena associated with a real-world domain that the work of art encodes. In the autoencoder case, we know to treat the artefactual collection of objects or phenomena as a systemic, structural representation of a domain whose aesthetic structure it idealizes: the manifold shape of a trained autoencoder's canon acts as a representation of the manifold-like shape of the training domain it idealizes, and provides a schema for interpreting the data in the training domain in accordance with a 'systemic gestalt' given by the manifold. Applying the same thinking to the literary case, we argue that a *dense aesthetic structure*[83] in the imaginative landscape associated with a work of art

[83]One hopes it isn't infelicitous to speak of a 'dense virtual, diffused, immanent' structure.

potentially acts as a representation of a *loose aesthetic structure* of the collective objects and phenomena of a real world domain. We argue, similarly, that the 'dense aesthetic structure' in question thus potentially provides a schema for interpreting the objects and phenomena of a real world domain in accordance with a 'systemic gestalt' given through the imaginative landscape of the literary work.

While it is true that, strictly speaking, an autoencoding manifold can always be directly mathematically expressed as the computational substrate of a trained autoencoder algorithm, the computations involved in autoencoding—let alone in any abstractly autoencoding-like bio-cognitive process—are mathematically intractable and conceptually oblique. If what we grasp in grasping the 'aesthetic unity' of some collection of phenomena is, at least in part, that this collection of phenomena can be approximated using a limited generative language, then we cannot hope to express or share what we grasped any more directly than by representing an idealized collection of phenomena. At the same time, this line of thinking promises that we can share our insight by intersubjectively constructing an appropriate set of idealized phenomena—and promises, as well, that the 'idea' that our set of idealized phenomena expresses is essentially impossible to paraphrase or separate from its expressive form, despite its worldly subject matter. An ambient meaning is therefore, in this sense, an *abstractum that cannot be separated from its concreta*. The above phrasing tellingly, if unintentionally, echoes and inverts a certain formula of the 'romantic theory of the symbol'—as given, for example, in Goethe's definition of a symbol as 'a living and momentary revelation of the inscrutable' in a particular, wherein 'the idea remains eternally and infinitely active and inaccessible [wirksam und unerreichbar] in the image, and even if expressed in all languages would still remain inexpressible [selbst in allen Sprachen ausgesprochen, doch unausppprechlich bliebe].' (Beistegui 24) The relationship of our literary-philosophical trope

of 'ambient meaning' to the romantic literary-philosophical trope of 'the Symbol' is even clearer when considering Yeats'ss more pithy paraphrase a century later, at the end of the romantic symbol's long trans-European journey from very early German romanticism to very late English Symbolism: 'A symbol is indeed the only possible expression of some invisible essence, a transparent lamp about a spiritual flame.' (25) A question therefore brings itself to mind: does the idea of an abstractum that cannot be separated from its concreta simply reaffirm the Goethe/Yeats theory of the symbol from the opposite direction, positing a type of abstractum (a 'structure of feelings') that can only be expressed in a particular, rather than a type of particular (a 'symbol') that singularly expresses an abstraction? Not quite, I would argue—and, indeed, I would suggest the difference between the two is key to the elective affinity between the theory of ambient meaning and specifically *Modernist* ars poetica.

While our study's working literary-historical framework treats certain historically Symbolist works as part and parcel of Modernist textual practice, and—crucially—treats Symbolist poetic theory as the 'mitochondrial Eve' of (Western) 20^{th} century aesthetic theory, I believe the era of Pound, Eliot, Joyce, and Stein is marked by the ascendency of a certain *materialist* reorientation of the Symbolist/romantic tradition. One relevant sense of 'materialist' is the sense that Daniel Albright explores in his study of Modernist poetic theory's borrowings from chemistry and physics, but a broader relevant sense of 'materialist' is closer to 'not-Platonist,' or to 'immanent' in the Deleuzian sense. Recalling Joyce's and Zukofsky's great allegiance to Aristotle, and perhaps observing that William Carlos Williams's 'no ideas but in things' is about as close as one can get to 'universalia in re' in English, we might usefully think of the 'horizontal' or materialist idealism we will here associate with Modernism as broadly *Aristotelian*. This broadly Aristotelian orientation has a literal dimension at the level of aesthetic theory, wherein the Modernist theorist tends to treat abstracta as *relations*

between concreta ('universalia in re') rather than treat abstracta as metaphysically terminal entities, as well as an 'analogical' dimension at the level of aesthetic practice, wherein the Modernist writer tends to be more interested in expressing an abstractum through many-to-many relations between concreta, where no single objects takes the role of a Platonic 'third man' that triangulates the multitude, than in expressing an abstractum through one-to-many or many-to-one relations between concreta.

For the Modernist aesthetic theorist, the philosophical burden on poetics partly shifts from the broadly Platonist burden of explaining how concreta could rise up to reach an otherwise inexpressible abstract idea, to the broadly Aristotelian burden of explaining how a set of concreta is (or can be) an abstract idea. Where Coleridge looked to the Imagination[84] as the faculty that vertically connects the world of things to the world of ideas, William Carlos Williams looked to the Imagination as the faculty that horizontally connects *things* to create a *world*. Similarly, where the early Schlegel of the 'Studium' mourns for the epochal loss of a once commonplace faculty for sensing transcendental truths, the early Eliot of 'The Metaphysical poets' instead mourns for the epochal loss of a once commonplace faculty that 'could devour any kind of experience.'[85] From a broadly Aristotelian point of view, the

[84]Coleridge and William Carlos Williams both take their concept of Imagination from Kant. It wouldn't be wholly ridiculous to say that Coleridge was concerned more with the 2nd critique and William Carlos Williams with the 1st.

[85]It is, of course, important to again stress that our gesture toward singling out Modernist practice or theory from its Symbolist or romantic roots as the paradigm example of a practice of ambient meaning, is at most a matter of Jakobsonian 'dominant' and elective affinity. It is far from impossible to find, among the German and British romantics, discourse on the Imagination as the synthesizing faculty that draws connections between the concreta of the world. Something akin to Eliot's view of the poetic faculty as a catalyst or medium that binds disparate materia together, for example, will sometimes appear in Coleridge (likely after Kant) in explicit form as a kind of first step in the Imagination's ladder from matter to meaning, which first synthesizes matter or sensation into a worldly coherence, then seeks a higher, supersensible truth by whose grace worldly coherence came to be. Goethe, whose practice in Faust II is paradigmatically Modernist in our sense, and whose ars poetica is strongly relational, nevertheless metaphysically separates meaning from matter. Baudelaire's 'correspondences' appear to vacillate between 'horizontal' and a 'vertical' conceptions.

Poundian/Eliotian—or, less canonically but more accurately, Steinian—operation wherein poetry explicitly arranges or aggregates objects in accordance with new, unfamiliar partitions[86] is precisely what it means to fully and directly represent abstracta: an abstractum *just is* the collective affinity of the objects in a class. In fact, in 'New Work for the Theory of Universals,' the premier contemporary scholastic materialist David Lewis formally proposes that universals are simply 'natural classes,' metaphysically identical to sets of objects that posses internal structural affinity.

By way of an example of a literary work's production of a 'horizontal' *symbol* as described above, we might observe that the imaginative landscape created by the collected writings of Franz Kafka plausibly functions as just this kind of aesthetic schema for unity or the affinity of a collection of real world phenomena. A reader of Kafka learns to see a kind of Kafkaesque aesthetic at play in the experience of going to the bank, in the experience of being broken-up with, in the experience of waking up in a daze, in the experience of being lost in a foreign city, or in the experience of a police interrogation, in part by learning that surprisingly many of the real life nuances of these experiences can be well-approximated in a literary world whose constructs are all fully bound to the aesthetic rules of Kafkaesque construction. We learn to grasp a Kafkaesque aesthetic logic in certain worldly phenomena, in other words, partly by learning that the pure Kafkaesque aesthetic logic of Kafka's literary world can generate a surprisingly good likeness of these worldly phenomena. This quick encounter with Kafka, and with the inevitable 'Kafkaesque,' also provides us with a good occasion to remark an interesting relationship between ambient meaning, literary polyvalence, and processes of concept-learning. Let us take the late French Symbolist and early Parisian avant-garde concept of 'polyvalence' to include both phenomena of collage, hybrid-

[86]A partition is the division of a set into non-overlapping subsets.

ity, and polyphony, where the heterogeneous multiplicity is on the page, and phenomena of indeterminacy, undecidability, and ambiguity where the heterogeneous multiplicity emerges in the readerly process. On the view suggested here, a vibe-coherent polyvalent literary object functions as a nearly-minimal concrete model of the abstract structure shared by the disparate experiences, objects, or phenomena spanned by the polyvalent object, allowing us to unify these various worldly phenomena under a predicate—e.g. the 'Kafkaesque.' The paradigmatic cases of this cognitive work are, inevitably, those that have rendered themselves invisible by their own thoroughness of impact, where the lexicalization of the aesthetically generated concept obscures the aesthetic process that constitutively underlies it: we effortlessly predicate a certain personal or institutional predicament as 'Kafkaesque,' a certain worldly conversation as 'Pinteresque,' a certain worldly puzzle as 'Borgesian.'

My near-repetition above of Shelley's complaint in the 'Defence' that all literal language is forgotten poetry warrants a consideration of the intellectual history of the relationship between literary polyvalence and concept-learning, which is something of a tortured subject. Outside of an arguable conceptual antecedent in Vico's theory or myth in 'New Science,' literary concept-making through polyvalence—the aggregation of heterogeneous materials—occupies a strange terrain between rarity and triviality in the history of artistic manifestos and aesthetic theory. It is almost tautological that Modernist and Symbolist and romantic theory espouse literature's work as revelation through polyphony and polyvalence, and 18$^{\text{th}}$ century philology has known a proliferation of theories of poetry as an originary language capable specifically of concept-making by the expressive power of meter and phonetics (e.g. Warburton, Herder, Rousseau),[87] but one is hard-pressed to find any explicit theories of *concept-making through polyphony and polyvalence*. While I believe that this paradigm makes

[87]See Peterfreund.

for a fruitful revisionary reading of some sections within Baumgarten's 'Aesthetica,' Shelley's 'Defence' stands as the one canonical text of Western aesthetics that makes an obvious explicit gesture towards the idea outside of Vico. Elsewhere, recently published correspondence between the central 20$^{\text{th}}$ century mathematician Andrej Kolmogorov—incidentally or not, the titular inventor of Kolmogorov complexity theory—and his colleagues in Yuri Lotman's Tartu school of semiotics takes up a related idea of literary works as 'standards of comparison' a subject may employ to structure her worldly gestalt. A view of the literary production of ambient meaning as a sometimes central pivot of societal concept-making processes positions literary polyvalence—though, admittedly, not *necessarily* in its stereotypically Modernist form—as crucial both to the construction of Jamesonian 'cognitive mapping' of the social and material world, and, complementarily, to the construction and transmission of ideology. As an approach to stereotypically Modernist polyvalence, our paradigm thus runs strongly counter to the common critical theory treatment of Modernist polyvalence as 'fragmentation,' contravening both its deflationary treatment within the discourse of Lukcs, Moretti, or Jameson and its lionizing treatment in the discourse of Adorno, Kristeva, or Barthes. In regarding Modernist polyvalence as a *method* of cognition rather than as a representation of cognition, one raises the possibility that the relationship of Modernist polyvalence/polyphony to the sociocultural and material conditions of modernity as flux, excess of information, and (prima facie) fragmentation of reality is strategic rather than metaphorical or expressive. On this epochal story, if we wish to tell it, the conditions of modernity prioritize continuous concept-learning as an essential cognitive survival tactic, or even as a favored strategy for cognitive thriving, much as concept-learning is the prioritized intellectual activity for a child encountering the world for the first time, and much as Alexander Grothendieck proposed that naming is the better part of mathematics.

3.3 System

One thing that makes the promise of a paradigm that treats a work of art as the representation of a worldly 'style' or 'vibe' in Ideal or idealized form promising, I would suggest, is that this paradigm organically calls on us to think of the 'aesthetic unity' of a work of art as a potentially world-directed property of the work—to think about it as, potentially, a part of what a work of art 'says' about the actual world. Perhaps the most intuitively powerful bridge between the raw 'aesthetic unity' of an autoencoding manifold and the kind of systemic modeling of reality that we associate with trained autoencoders is what we might call the relation of *comparability* between all objects in a trained autoencoder's manifold. The global aesthetic unity of the objects in a sensible autoencoding manifold, I want to argue, is not just technically but *conceptually* and *phenomenologically* inseparable from the global intercomparability of the manifold's objects, and the global intercomparability of the manifold's objects is not just technically but conceptually and phenomenologically inseparable from the representation of a *system*.

As we described in Chapter 1, the fact that the input-space points of a manifold map to a lower-dimensional space internal to the manifold means that whenever we compare two objects from the manifold, we can describe their difference as a line in the lower-dimensional space internal to the manifold instead of as a line in input-space. When our Classical Greek pottery experts from Chapter 1 compare two Classical Greek pots, for example, our experts can efficiently express the difference between the two pots in 'Classical Greek pot terms'—which our experts cannot do when they compare (e.g.) a Classical Greek pot and a dollar bill. While we have already discussed the technical grounds for this relation of comparability in the previous chapter, it's conceptually illuminating to consider how this comparability follows directly from the 'aesthetic unity' of the manifold's objects. Recall

again that an autoencoding manifold is, effectively, a collection of objects that aligns to a limited generative vocabulary. Because the objects that make up a trained autoencoder's manifold can all be specified using a single generative language, whenever we wish to compare two objects on the manifold we may, instead of comparing the 'visible properties' of the objects, compare the generative formulae that specify them. As we discussed, the generative language of an autoencoder has the structure of a space—specifically, the structure of a lower-dimensional space internal to an input-space manifold—which means that we can mark each generative formula as a list of numbers (coordinates in the lower-dimensional representation-space), and mark the relationship between any two generative formulae by subtracting one list from the other. What this means, in intuitive terms, is that the (lower-dimensional representation-space) difference between any two objects on the manifold is *itself* expressible in the autoencoder's 'language.'[88] The generative language that 'creates' the objects of the manifold is, in this sense, equally a 'transformation language.' We can, for example, re-imagine our Classical Greek pottery experts' experiment such that instead of sending Expert #2 the generative formula for each pot, Expert #1 sends Expert #2 the generative formula for the first pot in the exhibition, and then a series of *transformation formulae* that turn each pot into the next. For any pair of objects x, y on the manifold, the language of the trained autoencoder has a formula for transitioning from x to y. Conversely, no formula in the autoencoder's language, and no sequence of applications of formulae in the autoencoder's language, can transforms an object from the manifold into anything other than an object from the manifold. We might recall now that the manifold is, literally, the sum total of all possible movements along the *respects of variation* defined by the trained autoencoder's

[88]It's also possible, as logically follows, to treat the difference between any two points as itself the coordinates of a point, and so as a generative formula for an object that expresses their 'difference.' The conceptual robustness of this operation (in an ideal setting) is something of an open question, and in practice results have been mixed—sometimes striking, sometimes nonsensical—depending on the domain.

features. To stay on the manifold means to transform objects in the *same handful of ways*—strictly speaking, in different mixtures of the same handful of ways—again and again and again.

In the phenomenology of reading, we are experiencing this (so to speak) 'sameness of difference' as primary, and the 'aesthetic unity' of a literary work's imaginative landscape as derived. A literary work's 'style' or 'vibe,' is, at first, an invariant structure of the very transformations and transitions that make up the work's narrative and rhetorical *movement*. As we read Georg Büchner's 'Lenz,' for instance, plot moves, and the lyrical processes of Lenz's psyche revolve their gears, and Lenz shifts material and social sites, and every change consolidates and clarifies the higher-order constancy of mood. A given literary work's invariant style or vibe, we argued, is the aesthetic correlate of a literary work's internal *space of possibilities*. This space of possibilities is, from the reader's point of view, an extrapolation from the *space of transformations* that encodes the logic of the work's narrative, lyrical, and rhetorical 'difference engine.' Or, more prosaicly: to grasp a 'style' or 'vibe' should mean, no less than it means a capacity to judge whether a set of objects or phenomena does or does not collectively possess a given style, a capacity to judge the difference between two (style-conforming) objects in relation to its framework. Learning to sense a system, and learning to sense in relation to a system—learning to see a style, and learning to see in relation to a style—are, autoencoders or no autoencoders, more or less one and the same thing.[89] If the above is right, and an 'aesthetic unity' of the kind associated with a 'style' or 'vibe' is immediately a sensible representation of a logic of difference or change, functional access to the data-analysis capacities of a trained autoencoder's feature function and abstract

[89]I'd argue that the relationship between the two is weaker than a conceptual identity, but stronger than mere correlation. It is possible that it should best be understood as a Kripke-style necessary a posteriori identity, but it is plausibly a stronger relationship implicit in the phenomenology of style, and not only the cognitive mechanisms causally responsible for style-perception.

lower-dimensional representation-space follows, in the *very* long run, even from appropriate 'style perception' or 'vibe perception' alone, since the totality of representation-space distances between input-space points logically fixes the feature function. More practically, access to representation-space difference and even to representation-space distance alone is—if the representation-space is based upon a strong *lossy compression schema* for the domain—practicably sufficient for powerful 'transductive'[90] learning of concrete classification and prediction skills in the domain. When we grasp a loose 'vibe' of a real-life, worldly domain via its idealization as the 'style' or 'vibe' of an ambient literary work, then, we are plausibly doing at least as much 'cognitive mapping' as there is to found in the distance metric of a strong lossy compression schema.

One reason the mathematical-cognitive trope of autoencoding matters, I would argue, is that it describes the bare, first act of treating a collection of objects or phenomena as a set of *states of a system* rather than a bare collection of objects or phenomena—the minimal, ambient systematization that raises *stuff* to the level of *things*, raises *things* to the level of *world*, raises *one-thing-after-another* to the level of *experience*. (And, equally, the minimal, ambient systematization that erases nonconforming *stuff* on the authority of *things*, marginalizes nonconforming *things* to make a world, degenerates experience into false consciousness.)[91] In relating the input-space points of the manifold to points in the lower-dimensional internal space of the manifold, the manifold model creates the fundamental distinction between phenomena and noumena that turns the input-space points of the manifold into a system's range of visible states rather than a mere arbitrary set of phenomena. The parallel 'aesthetic

[90] See Vapnik.

[91] Cf. Adrian Piper's 'Xenophobia and Kantian Rationalism,' wherein Piper discusses xenophobia as 'a special case of a more general cognitive phenomenon, namely the disposition to resist the intrusion of anomalous data of any kind into a conceptual scheme whose internal rational coherence is necessary for preserving a unified and rationally integrated self.'

unity' in a world or in a work of art—what we have called a 'style' or 'vibe'—is arguably, in this sense, something like a maximally 'virtual' variant of Heideggerian mood. If a mood is a 'presumed view of the total picture' (Flatley) that conditions any specific attitude toward any particular thing, the aesthetic unity that associates the collected objects or phenomena of a world or work with a space of possibilities that gives its individual objects or phenomena meaning by relating them to a totality is sensible cognition of (something like) the Stimmung of a system—and much like Stimmung it's the 'precondition for, and medium of' all more specific operations of subjectivity; what an autoencoding gives is something like the system's basic system-hood.

On the view I'm suggesting here, the defining property of any system is the distinction between *visible states* and *underlying states*.[92] Much recent work in computational phonology, for example, has successfully induced autoencoding networks to take spoken English phonemes encoded as raw digital sound for input, and construct a manifold whose dimensions correspond to the fundamental dimensions of structural variation between English phonemes. Here, the input-space is the space of all possible (digital) short sounds, as represented by whatever general audio encoding format, and the lower-dimensional internal space of the manifold is the space of all spoken English phoneme sounds. We say that English phonemes are a system exactly because a spoken English phoneme sound can be described via an underlying phonetic structure that defines its position in the space of English phoneme sounds. From this perspective, the 'visible states' of the system of English phoneme sounds comprise the actual short sounds which are the input-space points of the manifold, and the 'underlying states' are the corresponding coordinates of these points in the lower-dimensional internal space of the manifold, where each English phoneme sound is a specific setting of the

[92]The Deleuzian reader may be interested in comparing the above with Deleuze's account of a 'structure' in 'What is Structuralism.'

machinery of fundamental phonetic structures. While generally we do not expect that the dimensions of a manifold model will correspond, as they sometimes have in computational phonology, to analytically discernible 'factors of variation,' the analytically ineffable factors of variation associated with the dimensions of a manifold that models a more worldly, messy domain turn the collection of phenomena associated with the domain into a *system* in just the same way, by mapping variation in visible phenomena to changes in the state of an underlying system.

To see what we can make of the above in literary terms, we will now take a short walk through a light, provisional 'ambient-systemic' reading of John Ashbery's iconic short lyric 'At North Farm.'

> Somewhere someone is traveling furiously toward you,
> At incredible speed, traveling day and night,
> Through blizzards and desert heat, across torrents, through narrow passes.
> But will he know where to find you,
> Recognize you when he sees you,
> Give you the thing he has for you?
>
> Hardly anything grows here,
> Yet the granaries are bursting with meal,
> The sacks of meal piled to the rafters.
> The streams run with sweetness, fattening fish;
> Birds darken the sky. Is it enough
> That the dish of milk is set out at night,
> That we think of him sometimes,
> Sometimes and always, with mixed feelings?

I want to read Ashbery's prosperous and gloomy[93] 'At North Farm' in terms of two central 'moments': the construction of a vibe, and the intellection of a worldly system in the recognition of this vibe. Parts of the reading will be unabashedly structuralist in flavor, with

[93]The name and driving image of the poem come, as Ashbery explains, from the Finnish epic 'The Kalevala': 'North Farm in the epic is a place near hell but not in it, and it is always referred to with the epithet 'gloomy and prosperous North Farm." (Lehman)

the hope that one accepts that the structural analysis is given as critical heuristic, not as literary, aesthetic, or computational noumena. (Recall that on a neural network framework we are always crudely approximating a process too 'soft' to correspond to any description in English.) The first moment of the text is constructed by inverting the cosmological schema of 'Sailing to Byzantium' via Kafka's 'A Message from the Emperor.' Recall that 'A Message from the Emperor' is a journey beginning in an emperor's throne-room, and 'Sailing to Byzantium' a journey ending in an emperor's throne-room. In 'A Message from the Emperor,' the transcendent inner sanctum (the 'imperial sun') is a deathbed and a runner's start-line. Contra Yeats's Byzantium, its remove from the world makes it the one place where things can be begotten, born, and die. On entering the world the energy of the process originating in the transcendence of the imperial sun is dampened into stasis by the density of the crowd and of the palatial and urban infrastructure, and the world's imperviousness to the temporal continuously intensifies as the narration draws towards the workaday reality of the city where 'you' live and daydream at your windowsill. Ashbery substitutes for Kafka's replete streets and courtyards the repleteness of Yeats's 'salmon falls' and 'mackerel-crowded seas,' transforming the image of earthly plenty from a *figura* of birth-and-death to the *figura* of a dis/utopian world-system perfected in its artifice: 'Hardly anything grows here,/ Yet the granaries are bursting with meal,/ The sacks of meal piled to the rafters./The streams run with sweetness, fattening fish; /Birds darken the sky.'

The land in 'At North Farm' inherits the sensual repleteness of the Yeatsian source, but reverses its association with the natural and with the transitory. On the resultant inversion of 'Sailing to Byzantium,' the transient world of earthly immanence is turned timeless, the world of the transcendent is turned temporal, the voyager is turned into a host, nature is turned into culture, the quest is turned into a daily ritual, and the origin is turned into a

destination. Thus in 'At North Farm' the world of earthly plenty—of Yeats's 'fish, flesh, or fowl,' 'birds in the trees' and 'salmon falls' and 'mackerel-crowded seas,' or Ashbery's 'sacks of meal piled to the rafters,' 'streams [running] with sweetness, fattening fish' and 'birds [darkening] the sky'—is a world without time or becoming where 'hardly anything grows,' suggesting that there's nothing much around that is 'begotten, born and dies' as things do in the flux of Yeats's earthly world's self-commendations. In breaking Yeats's pairing of the earthly with the natural and the transitory and of the transcendent with the cultured and the timeless, Ashbery takes apart the Apollonian/Dionysian dichotomy of Yeats's schema and opens for consideration the nature of the conceptual space (a space of possible chronotopes, one might say) that emerges in the combinatorial play of tropes of time, timelessness, earthly plenty, transcendent beyond, nature, culture, agency, passivity, mortality, temporal distance, and physical distance—a conceptual space in which the Apollonian/Dionysian dichotomy of the 'Sailing to Byzantium' chronotope is a particular coordinate.

In at least a weak sense, 'At North Farm' imputes a ***parameterization*** for the space around 'Sailing to Bynzantium,' defining the respects in which something can be like or unlike it. The space is initially 'book-ended' by 'Sailing to Byzantium' and 'A Letter to the Emperor,' with 'A Letter to the Emperor' marking an *opposite setting* of all the parameters that define 'Sailing to Byzantium.' Each further complication of the cosmological schema through analogical mapping to another work of literature either adds another dimension to the feature-space, or leads to a more abstract reformulation of the existing features. Let us be as theoretically clear as we can: I am surmising that the construction of the feature-space is guided by an imperative to preserve the information contained in the schemas induced by each analogical mapping, so that each mapping that produces a new schema can only add dimensions to the feature-space. At the same time, one wants to keep the number of

dimensions as low as possible, so one seeks to factorize the aggregated schemas into more basic respects of variation. To see an example at play, let us focus more deeply on the relationship between mass,[94] movement, and time, this time considering the parametrization of a manifold that also encompasses Milton's 'On his Blindness' in addition to 'At North Farm,' 'Sailing to Byzantium,' and 'A Letter to the Emperor.'

Recall that the account of the repleteness of the land in 'At North Farm' is given a parallel position (in terms of the rhetorical flow) to the account of the density of crowds and buildings in 'A Message from the Emperor': the description of the repleteness of the land comes after, and in juxtaposition to, the description of the speed of the traveler. The description of the repleteness of the city similarly comes after, and in juxtaposition to, the description of the speed of the messenger. This parallelism suggests expanding the opposition between Kafka's crowds and narrative time to a more abstract opposition between spatial density as such and temporality as such. That is, we want an opposition that enfolds both the opposition between spatial density and temporality in Kafka's story and Ashbery's opposition in 'At North Farm' between the repleteness of the Yeatsian plenty and the temporality of growth/labor and of narrative. This mapping between 'At North Farm' and 'A Message from the Emperor' induces the schema of a 'mass vs. time' chronotope of which both works can now be considered exemplars. The induction of this chronotope arguably gets at the root of the crowd's role in 'A Message from the Emperor,' allowing for a cognitive representation that is not just general but indeed 'deep': 'A Message from the Emperor' would not have been an effective work of art if the connection between crowdedness and timelessness relied solely on the literal role of the crowd as a physical roadblock. If we read 'A Message from the Emperor' alone, a natural first-step symbolic interpretation of the crowd's anti-narrative effect would

[94]By 'mass' we mean a mass of things or of stuff, not the physical-scientific quantity.

relate the crowd's disruption of narrative process to a tension, typically associated with modernity, between the crowd as a social mass and the temporality of the stereotypically bourgeois or monolithic subjectivity associated with the rise of the Bildungsroman. By contrast, the schema induced by mapping 'A Message from the Emperor' to 'At North Farm' suggests abstracting further to arrive at a chronotope defined by an antinomy between the phenomenology of spatial density and the phenomenology of temporality. (One might even think here of Heidegger's opposition of 'fascination' and 'being-towards-death.') This is roughly the opposite chronotope to the chronotope of the Yeatsian juxtaposition between an avalanche of mortal transient rejoicing bodies and a timeless ornament at an emperor's throne.

The feature-space defining the manifold of chronotopes is further refined by the activity of mapping/comparing Milton's 'On His Blindness' to the resultant schema, which breaks apart the 'temporality' role from the earlier schema by pairing one aspect of narrative time—action—with mass ('*thousands* at his bidding speed...'), and another aspect of narrative time—expectation—against it. The relationship of 'At North Farm' to 'On His Blindness' is first established in the description of movement in the third line of 'At North Farm': the line 'traveling day and night, through blizzards and desert heat, across torrents' invokes Milton's 'post o're Land and Ocean without rest,' and so positions 'On His Blindness' as a second near-twin to 'At North Farm,' alongside 'A Message from the Emperor.' Yet the crucial activation of the intertext takes place when we map 'On His Blindness' to 'A Message to the Emperor,' as 'On His Blindness' melds together the crowd and the messenger of 'A Message from the Emperor': '*Thousands* at his bidding speed and post o're Land and Ocean *without rest*.' This complication breaks with the mass vs. narrative-time chronotope shared between 'At North Farm' and 'A Message from the Emperor' by splitting the alignment of two aspects

of narrative time, namely agency and being-towards-death/expectation, and pairing mass with agency but opposite being-towards-death/expectation. This change requires breaking up the 'narrative time' role in the schema into two separate roles, since in 'On His Blindness' the two aspects of narrative time have a different relationship to mass, which the schema produced by the earlier mapping is obligated to record.

As technical-minded as this reading has been, I believe this first moment of Ashbery's text corresponds to an experience ('vibe') of the feature-space as a 'manifold of immanence' that is held constant across the different settings of the parameters. This is not an experience of the different structures, poems, and exemplars as being alike, but rather the experience of moving around a sensibly coherent plane in moving between them. The feature-space of 'At North Farm' could perhaps be called the world-system of petite bourgeoisie eschatology or 'micro' eschatology. The second moment of 'At North Farm,' dependent on the first, considers the relationship between the feature-space/manifold to points *outside* the manifold. These relations arise through the polyvalent metonymic activation of tropes in the 'tubular neighborhood' of North Farm's manifold—that is, tropes have a reasonable projection distance to the manifold, whereby it's plausible that these tropes are structurally elucidated, as much as effaced, by their reduction in accordance with North Farm's lossy compression schema. From this point of view, we can consider 'At North Farm' as a manifold whose tubular region encompasses Yeats's 'Sailing to Byzantium,' Kafka's 'A Message from the Emperor', the trope of waiting for Santa Claus (cf. leaving out milk) as a child (cf. having no agency and having your material needs super provided for in mysterious ways), the trope of waiting for love or religion or revolution as an adult living in a city at the center of an international capitalist world-system (cf. having no agency and having your material needs super provided for in mysterious ways), and the memento mori of the 'vanitas' tradition

(cf. grim ripper figures you are hiding from amidst the worldly riches). At the same time, projection to the 'North Farm' manifold comes close to nullifying the distinction between the habitual and tropes of the prolepticly eschatological, allowing a glass of milk to polyvalently drag together leaving out some milk for the street cats that come by at night or for your house cat, leaving a glass of milk to Santa Claus, and leaving out food to appease spirits that might come during the night with little dissonance or contrast, and perhaps no great loss of meaning—a provisional cognitive-aesthetic demonstrating that things done for the sake of some eschatological hope or fear are, in a good deal of respects, not very different from normal minor daily habits.

3.4 Structures of What

As my reader may have noticed, the above does not say much about how we should understand the content of ambient meaning proper, prior to its more familiar partial interpretations as a mood, as a style/vibe, or as system. The content of ambient meaning, I will argue, both is and is not a thing apart from these three partial interpretations. I identify ambient meaning, in the last instance, with the epistemic form of those ambiguously social, psychical, and cultural-material structures that Raymond Williams called '*structures of feeling*': structures waiting 'at the very edge of semantic availability,' each 'a quality... which gives the sense of a generation or a period.' (131) While critics sometimes group 'structures of feelings' with more specifically defined structures like Heidegger's 'mood' or Bourdieu's habitus, I understand Williams's structures not as moods but as something closer to a limit-case of systemic structure in general—systemic structure operating, as Williams would have it, 'at the very edge of semantic availability.' I propose that ambient meaning, as a mapping that abstracts beyond the distinction between mood, system, and style/vibe, is best understood as a di-

rect representation of such a Williamsian 'structures of feeling'—a direct representation of a cultural-materialist version of a 'ghostly paradigm of things,' if you will. At the same time, I argue that because of the deep ontological remove of a 'structure of feeling' from any direct concrete matter, we can only verbalize a structure of feeling in relation to the more concrete (if epistemically elusive) trio of mood, system, and style/vibe—a limitation that makes a 'three-headed' engagement with ambient meaning as mood, system, and a style/vibe indispensable. The strong complementarity between ambient meaning and Williams's 'structures of feeling' is perhaps best seen by delving into Williams's systematic use of indeterminacy—that is to say, deep verbal or conceptual ambiguity—to coin the concept in 1977's 'Marxism and Literature.' Formally defining 'structure of feeling' as 'a particular quality of social experience and relationship... which gives the sense of a generation or a period,' Williams makes no move to settle whether 'feeling' comes into the term in reference to the 'social experience and relationships' part of the definition, as synonym for 'experience,' or comes in reference to the 'sense of a generation or a period' part of the definition, as a synonym for 'sense.' Neither does Williams move to settle whether the 'sense of a generation or period' refers to a generation or period's historical-materialist meaning or raison d'être, or to a historian's sense that some social-material corpus constitutes a generation or a period, or to a generation or period's native sensibility or sense of meaning. Nor, finally, is there discerning whether the connective 'of' ('structure of feeling') should mean structure made of feeling, or a structure that regulates feeling, or a structure that is felt.

Rather than indicating any surfeit of expository patience, the extreme ambiguity of Williams's definition mirrors the ontology-defying scope of the problem Williams sets for his 'structures of feeling,' which are meant to traverse 'all that escapes from the fixed and explicit and the known.' In Williams's practice, the extreme ambiguity of definition turns

into an extreme indeterminacy or abstraction in the nature of the concept. Williams variably describes the problem of 'structure of feeling' as the problem of accounting for the role of sensible knowledge in social action, as the problem of emphatically accessing affective lives within the framework of a structural understanding of the social, and as the problem of articulating system-level social-material patterns too diffuse for the historian to articulate explicitly, never quite suggesting that these various descriptions (manifestly different in their respectively social-affective, cognitive-aesthetic, and social-material orientation) aren't interchangeable synonyms. The challenge Williams is broaching, I think, is the problem of defining an ambient 'thing-in-itself,' beyond the ontological separation between mood, style/vibe, and system. I propose that we might try identifying such a structure with a kind of 'metastable fixed point' of a world defined by the reciprocal dynamics of mood, style/vibe, and system. The 'very edge of semantic availability' being what it is, however, let us observe that by contrast with the committed application of mathematical concepts and arguments in the mainstay of this dissertation, the broadly 'systems ecological' social-aesthetic story I propose below is strictly in the realm of myth or metaphor.

On the picture that I am suggesting, there exists a *reciprocity* between the structure of our sensibility or sensible cognition (system), the structure of our affective life or social experience (mood), and the structure of our social-material performance or production (style/vibe)—a reciprocity whose approximate equilibrium or 'metastable state' binds the cognitive, affective, and material aspects of life into a coherent lifeworld or 'totality.' One way to tell the story of this reciprocity is as follows. The system of our sensibility—our faculty of sensuous cognition that discloses objects, properties, and patterns—recapitulates the structure of the social-material world. We continuously calibrate our sensibility by attuning it to our social-material world's dominant patterns and forms, adapting our powers of apperception

to the task of navigating our social-material world. This social-material world, however, is in no way an unmoved mover unilaterally structuring our affective and cognitive life: as a sum of human artifacts and social performances, the social-material world depends on human agency and on the subjectivity that animates it. In this regard, the structures, forms, and patterns of the social-material world do not just condition our sensibility, but also bear the imprints of our affective and cognitive life. One aspect of this social-material world in particular, which we may call the social-material world's (from the point of view of social-material production) style/vibe, is sensitive to our subjective life to such a point that it could be considered an expression of the subjectivity that animates our social-material production. A culture's style/vibe or feel, on this account, is the expression of a structure of the social-affective life (the feelings, drives, affects, imaginations) of the subjects whose collective social performances and cultural productions constitute the social-material world. According to this view one recognizes, so to speak, the 'touch' of a mood in the textures of the cultural-material production that it animates: the mood underlying our cultural-material production manifests as diffused textural affinity or formal constancy across the artifacts and social performances that constitute our social-material product, which we then experience as social-material 'style/vibe.' A social-material world's 'style/vibe' is thus a kind structural *family-resemblance* or *cognateness* between social-material products, rooted in a common social-affective origin.

So far, then, we have proposed a kind of chain of recapitulation leading us from system to style/vibe to mood, with style/vibe acting as the middle term. Style/vibe, we have suggested, recapitulates the structure of our social-affective world within the patterns and forms of the social-material world, and by these same patterns and forms determines the attunement of our system. We can—and in fact, should—now let this chain become a circle,

wherein every term acts as a 'middle term' that recapitulates one structure while determining the other, by subjecting mood to the same analysis that we applied to style/vibe. Mood, we might say, determines style/vibe but recapitulates our cognitive-aesthetic system. In a cognitive-aesthetic social analysis of the kind that Rancire famously practiced in 'The Distribution of the Sensible,' for example, we treat the virtual, diffused, immanent structure reigning over our affects, experiences, drives, and imaginations as a kind of family-resemblance or cognateness between feelings that developed in the bounds of the same social cognitive-aesthetic regime. The calibration of our cognitive-aesthetic system defines the limits of thought and perception ('distribution of the sensible') within which our drives, affects, experiences, and imagination find their catalysts and objects, literally structuring the space of our affective life. This, then, is one story of our cycle of reciprocation. Our cognitive-aesthetic system is calibrated to the structures of the social-material world, not least among them our social-material world's style/vibe. Our social-material world's style/vibe, in turn, expresses the mood animating our cultural-material production. Our mood, finally, is an extension of the 'distribution of the sensible' set by our social cognitive-aesthetic system.

I propose that we identify Williams's 'structure of feeling' with the above process of reciprocation operating at a *metastable fixed point*.[95] A Williamsian 'structure of feeling,' on this view, is the dynamics of the social-cultural ecosystem over a specific span—a 'generation or a period'—where the transmission from (e.g.) a style to system to mood to style reproduces style with *relatively little change*. Recall that Williams defines 'structure of feeling' as 'a quality... that gives the sense of a generation or a period.' Given the account of social-

[95]'Metastable'—a much plainer concept than the 'meta' would suggest—means slowly moving toward instability but 'stable for now.' A 'fixed point,' for a system that takes its own output as feedback, is an input that induces an identical output. A 'metastable fixed point' is, accordingly, roughly an input that nearly reproduces itself, and continues to change only slowly over many feedback iterations until reaching a breaking point wherein the next input and the next output are no longer close to one another.

cultural reciprocity we have proposed, a 'generation or a period' in Williams's sense should last exactly for however few or many iterations of the cycle of reciprocation the ecology of cognitive-aesthetic system, style/vibe, and mood stays within a given metastable state. But is this 'metastable state' the kind of thing that we should, or can, call a 'structure'? If this metastable state that makes a lifeworld out of our cognitive-aesthetic, social-affective, and social-material lives could be cognized or represented as a structure in any meaningful sense, then it would surely qualify as the Williamsian 'structure of feeling' par excellence: an abstract pivot of 'all that escapes from the fixed and explicit and the known,' which is at once the quality that makes a generation or a period its distinct self, a period's essential pattern of affective life, a period's mode of sensibility, and the characteristic vibe of its performances and artifacts.

This question of cognizing our ecological 'stable state' as a structure therefore brings us, finally, back to ambient meaning. Ambient meaning re-imagines the reciprocal relationship between system, mood, and style/vibe as a relationship of each to an Ideal structural common cause.[96] In an ambient representation, the structure of a mood, the structure of a style/vibe, and the structure of a cognitive-aesthetic system are all *applications* of the structure of an autoencoding manifold, whose multiplicity of meaning is only possible on account of the metastable fixed point. Or—

> 'How exquisitely the individual Mind
> (And the progressive powers perhaps no less
> Of the whole species) to the external World
> Is fitted:—and how exquisitely, too—
> Theme this but little heard of among men—
> The external World is fitted to the Mind; And the creation (by no lower name
> Can it be called) which they with blended might Accomplish'

[96]This theoretical fiction is permitted by the same logic by which physicists would introduce 'attractors' to a model of a dynamical system.

Works Cited