UTKU B. DEMIR

NOMADIC DESCENT:
GENERATIVE AI, SUBJECTIVATION, AND RESISTANCE/CRITIQUE IN CONTROL SOCIETIES

Abstract

The thesis examines how critique and resistance can be re-theorised in the context of the rapidly growing presence of Generative Artificial Intelligence (genAI) by first situating these systems within Gilles Deleuze's account of control societies. In this framework, classical institutions give way to computationally reinforced infrastructures that operate through personalised, flexible, and continuous modulation, reshaping the field in which subjectivation unfolds. Earlier AI systems have already shown a striking resemblance to the formulation of control through predictive relevance assignment and behavioural personalisation; contemporary genAI models go a step further with their novel capabilities and actively participate in the production of knowledge, rendering them important agents in the formation of human subjectivity. After a theoretical, historical, and technical analysis, the thesis also examines central contemporary debates around genAI, interrogating the nature of knowledge production in transformer-based architectures, the conditions of human–machine interaction, the reconfiguration of agency, and competing visions of model development. Drawing on Gilles Deleuze and Félix Guattari's project Capitalism and Schizophrenia, it mobilises concepts such as desiring-production, schizoanalysis, and nomadology to build a theoretical scaffold for rethinking how generative infrastructures and human–machine interactions might be shaped into divergent, non-sedimentary formations. Combining this framework with experimental interventions into model behaviour, the study argues that possibilities for critique and resistance emerge immanently within generative systems and their communicative dynamics. With the display of interventions like weight amplification, artificial curiosity, and counter-sequencing, the thesis demonstrates how generative dispositifs can be repurposed to activate divergent potentials, offering a micropolitical framework for critique and resistance.

Abstract (Deutsch)

Die vorliegende Arbeit untersucht, wie Kritik und Widerstand im Kontext der rasant zunehmenden Präsenz von Generativer Künstlicher Intelligenz (genAI) neu theoretisiert werden können, indem diese Systeme zunächst innerhalb von Gilles Deleuze' Konzept der Kontrollgesellschaften verortet werden. In diesem Rahmen treten klassische Institutionen zugunsten rechnergestützter Infrastrukturen zurück, die durch personalisierte, flexible und kontinuierliche Modulation operieren und so das Feld neu gestalten, in dem sich Prozesse der Subjektivierung entfalten. Frühere KI-Systeme zeigten bereits eine auffällige Ähnlichkeit zur Formulierung von Kontrolle durch prädiktive Relevanzzuweisung und verhaltensbezogene Personalisierung; zeitgenössische genAI-Modelle gehen mit ihren neuartigen Fähigkeiten jedoch einen Schritt weiter und nehmen aktiv an der Wissensproduktion teil, wodurch sie zu zentralen Akteuren in der Bildung menschlicher Subjektivität werden. Auf Grundlage einer theoretischen, historischen und technischen Analyse beleuchtet die Arbeit anschließend zentrale aktuelle Debatten um genAI, untersucht die Bedingungen der Wissensproduktion in Transformer-Architekturen, die Dynamiken der Mensch–Maschine-Interaktion, die Neukonfiguration von Handlungsfähigkeit sowie konkurrierende Entwicklungsparadigmen solcher Modelle. Unter Rückgriff auf Gilles Deleuze' und Félix Guattaris Projekt "Kapitalismus und Schizophrenie" mobilisiert die Arbeit Konzepte wie Wunschproduktion, Schizoanalyse und Nomadologie, um ein theoretisches Gerüst zu entwickeln, das neu denken lässt, wie generative Infrastrukturen und Mensch–Maschine-Relationen in divergente, nicht-sedimentierte Formationen überführt werden können. In Kombination mit experimentellen Eingriffen in das Modellverhalten argumentiert die Studie, dass Möglichkeiten für Kritik und Widerstand immanent innerhalb generativer Systeme und ihrer kommunikativen Dynamiken entstehen. Anhand von Interventionen wie Gewichtsverstärkung, künstlicher Neugier und Gegen-Sequenzierung zeigt die Arbeit, wie sich generative Dispositive umnutzen lassen, um divergente Potenziale zu aktivieren, und entwickelt damit ein mikropolitisches Rahmenkonzept für Kritik und Widerstand.

Acknowledgements

This thesis owes special thanks to my supervisor Sergej Seitz for his continuous support, patience, and incisive insights throughout the research and writing process. I am also deeply grateful to Christoph Hubatschke for generously taking on the role of co-supervisor and for his critical feedback and inspiring ideas. Finally, I thank Fabio Wolkenstein for his assistance in the early conceptualisation of the research topic and for his valuable guidance.

To my son Jonas, and my sister Bahar...

Contents

List of Figures

Glossary

assemblage A concept developed by Deleuze and Guattari to describe heterogeneous constellations of material elements (bodies, objects, infrastructures) and discursive components (rules, practices, ideas) that function together without forming a unified whole. An example derived from Foucault's analysis is the prison, which combines architecture, guards, inmates, and routines with legal codes and discourses on criminality. Assemblages are contingent, evolving configurations that integrate but do not fuse their parts, and they provide the conditions through which desire and social organisation manifest (see Buchanan 2018, 66-67). 14, 23, 35, 38, 70, 71, 74, 76, 96, 97

dispositif A dispositif (often translated as "device," "deployment," "apparatus," or "setup," though sometimes left untranslated) is, in Foucault's usage, elements of a heterogeneous network of discourses, institutions, architectural arrangements, regulatory rules, technologies, and practices. Rather than locating power in a single structure or person, the dispositif describes how power operates through relations and resistances embedded in everyday formations. It is central to Foucault's analyses of genealogies, biopower, and governmentality (see Crano 2020). 11, 12, 13, 14, 18, 19, 20, 21, 25, 28, 30, 33, 35, 36, 38, 47, 63, 68, 83, 84, 85, 86, 88, 101, 102, 107, 108, 110, 111, 113, 114

epoch Epochs represent the number of times the entire training dataset passed through the algorithm (Nebius-Team 2024). 55, 56, 60, 61, 67

kernel In Machine Learning, the Kernel method consists of using a linear classifier to solve a non-linear problem. This is achieved by transforming a linearly inseparable set of data into a linearly separable set (Melanie 2024). 51

loss/cost function A mathematical rule that quantifies the difference between a model's prediction and the correct outcome in order to *punish* weak predictions and *reward* the correct ones in following processes. For example, if the model predicts "cat" but the true label is "dog," the loss is high; if it predicts "cat" and the true label is also "cat," the loss is low. Training minimises this loss so the model improves (see Goodfellow et al. 2016, 178). 55, 56

neuron An artificial neuron is the basic building block of NNs, inspired by how biological neurons work. It takes several inputs, multiplies each by a weight,

adds them together with a bias, and then passes the result through an activation function to decide the output. This simple mechanism allows networks of many neurons to learn patterns and make complex decisions (see McCulloch and Pitts 1943 for the early initiation of neurons). 47, 57, 58, 98

token Tokens are the smallest units of text that a model processes; typically words, subwords, or characters in natural language processing tasks (Cser 2024). In current AI systems, a token often corresponds to a single word, and the process of breaking text into tokens is known as tokenization (Jurafsky et al. 2009, 59). For genAI models, particulary LLMs, this enables efficient computation across varying text inputs. 45, 49, 51, 52, 53, 55, 60, 65

Acronyms

AGI Artificial General Intelligence. 24, 65

AI Artificial Intelligence. 3, 4, 5, 10, 11, 13, 16, 17, 18, 20, 23, 24, 36, 38, 40, 41, 42, 43, 44, 45, 47, 55, 57, 61, 63, 64, 65, 66, 72, 74, 75, 76, 77, 78, 79, 80, 81, 85, 86, 90, 92, 94, 101, 106, 107, 109, 110, 112, 113

BwO Body without Organs. 59, 60

CNN Convolutional Neural Network. 49, 51, 103

CoPhe Computational Phenomenology. 4, 76, 78, 79, 80, 99, 111, 112

D&G Gilles Deleuze & Felix Guattari. 11, 14, 19, 20, 27, 32, 34, 42, 43, 54, 59, 60, 70, 81, 83, 84, 85, 86, 88, 91, 93, 95, 96, 97, 99, 102, 107, 108, 112

DL Deep Learning. 3, 20, 41, 44, 47, 48, 49, 55, 57, 59, 61, 68, 69, 72, 77, 79, 99, 110, 111

DNN Deep Artificial Neural Network. 10, 20, 23, 44

GAN General Adversarial Network. 106

genAI Generative Artificial Intelligence. 3, 4, 7, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 33, 37, 38, 40, 41, 43, 45, 47, 48, 49, 51, 52, 53, 59, 60, 61, 62, 63, 64, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 76, 78, 79, 80, 81, 85, 86, 87, 88, 89, 90, 91, 92, 93, 96, 97, 99, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114

GOFAI Good old-fashioned AI. 41, 57

LLM Large Language Model. 5, 7, 15, 16, 20, 23, 24, 33, 40, 41, 48, 49, 51, 52, 53, 56, 60, 61, 63, 64, 65, 66, 67, 71, 72, 74, 78, 89, 90, 91, 92, 96, 97, 99, 101, 106, 110, 111

ML Machine Learning. 41, 44, 49, 55, 58, 68

NLP Natural Language Processing. 3, 10, 23, 41, 43, 44, 49, 52, 54

NN Artificial Neural Network. 5, 6, 10, 17, 20, 23, 40, 41, 43, 44, 45, 47, 48, 49, 52, 55, 56, 57, 58, 59, 60, 61, 65, 68, 71, 76, 77, 78, 92

NR Neuro-Representationalism. 4, 76, 77, 80, 111

PCA Principal Component Analysis. 49, 50, 69

RLHF Reinforcement Learning from Human Feedback. 60, 61, 90, 93, 98, 102

RNN Recurrent Neural Network. 49, 51

SL Supervised Learning. 44, 66, 90

SSL Self-Supervised Learning. 20, 41, 44, 45

symAI Symbolic Artificial Intelligence. 20, 41, 42, 57, 110

UL Unsupervised Learning. 44, 66, 90

1

Introduction

The rapid integration of advanced Artificial Intelligence (AI) into the fabric of social life signals a profound shift in how power is organised and exercised. This transition finds a powerful precursor in Gilles Deleuze's (1992) short but influential essay "Postscript on the Societies of Control", where he identifies a move away from traditional, enclosed institutional formations towards a more fluid and pervasive form of power. In control societies, power is enacted through digital and computational infrastructures that operate by circulating data, generating predictions, and shaping behaviour in real time. Such infrastructures no longer simply regulate individuals from the outside; they participate directly in producing the very conditions under which subjects perceive, decide, and act. It is within this transformation that the present study situates itself, examining how contemporary Generative Artificial Intelligence (genAI) systems extend this machinery by functioning as meaning-making entities reshaping the conditions under which human subjectivity is formed.

Recent years have witnessed substantial advancements regarding computational infrastructures, especially in the field of AI. Innovations in Artificial Neural Networks (NNs) and Deep Artificial Neural Networks (DNNs) have enabled increasingly sophisticated solutions to Natural Language Processing (NLP) tasks such as text classification and topic modelling, with applications spanning search engines, social media feeds, streaming platforms, employment procedures, and surveillance infrastructures. These applications were largely predictive, focused on analysis, relevance association, and personalisation, resembling the machinery Deleuze has outlined. Yet, recently a new paradigm has emerged in the form of genAI. What began in the 1950s (Cao et al. 2023, 4) as a relatively marginal pursuit within the field of NLP, now underpins systems capable of generating novel text, images, and code, translating human prompts into coherent outputs by drawing on vast reservoirs of raw data. Far from being mere technical novelties, and unlike earlier models limited to prediction, they have become cultural phenomena. Models like ChatGPT, Stable Diffusion, and other successors are now active participants in knowledge production, communication, and cultural creation.

The advancements signal a fundamental reconfiguration of AI's epistemic role and operational qualities: from predictive instruments to generative infrastructures that govern and (re-)produce knowledge itself. By functioning as meaning-making entities (Dishon 2024), genAI systems not only reshape institutional pro-

cesses central to shaping human subjectivity (MacKenzie and Porter 2021), but also fundamentally challenge established sociotechnological imaginaries of artificial agents. Particularly, their *transformer-based* architectures and novel *attention* mechanisms enable genAI models to draw connections across widely separated linguistic and visual patterns (see Montanari 2025). This capacity allows them to generate coherent outputs that increasingly mediate social reality and shape the interpretative frameworks through which subjects navigate the world. At stake is not only how such systems generate information and reorganise meaning, rather a decisive evolution, one where power operates not by shaping what subjects see, but by generating the very fabric of what can be seen and thought, thereby challenging the possibility of critique at its source.

It is precisely this enclosure of the imaginative and interpretive terrain that demands a re-theorisation of resistance. Yet, the critical theory of AI currently lacks a robust formulation of critique and resistance that is grounded in a technical analysis of these systems. In this thesis, I build on critical perspectives from political theory and the philosophy of technology to interrogate the institutional, epistemological, and political implications of contemporary genAI systems by analysing their architectural structures in depth. I aim to address the lack by formulating a theory of resistance that works through and with these generative infrastructures to counter or divert specific tendencies in processes of subjectification central to the machinery of control. The task is tackled by pursuing three main prospects after situating genAI within Deleuze's control society: a technical analysis of the mechanisms giving life to these architectures, a discussion of the most prominent debates around genAI, and an articulation through Gilles Deleuze & Felix Guattari (D&G)'s broader project "Capitalism & Schizophrenia" (see 1983 and 1987) and its unique contribution to revolutionary theory.

1.1 *Charting a Manifold: Research Question & Motivation*

One could argue that AI is "no longer an engineering discipline" (Dignum 2023, 206), if indeed it ever truly was. Each advancement in the design of systems that transform data[1] into interpretations of the world simultaneously reconfigures relations of power and knowledge. Algorithmic capacities for decision making, information management, content creation, and narrativisation turn AI systems into political entities. Their nature, formation, and the functions they perform must therefore be examined as parts in the machinery of power, or in a more comprehensive sense as dispositifs: dynamic arrangements through which technical architectures, institutional practices, and epistemic frameworks are articulated to produce, distribute, and regulate knowledge. To look under the hood of these sociotechnological dispositifs is to discover less about their internal cogs and gears than about the power structures they reproduce and sustain. What matters, then, is not simply the technical functioning of AI but the ways in which it embeds itself into everyday life, binding knowledge to governance and influencing how individuals come to understand themselves and their world. This move requires a conceptual vocabulary that accounts for the imbrication of knowledge and power, and Foucault's formulation of the Power/Knowledge nexus provides precisely such an entry point (Foucault 1980, 109–134).

[1] Read: traces of the human past. See Denton et al. (2021) for a detailed analysis of datasets, Jones (2023) for a concise critical account of their role, and Jones (2025) for a more extensive treatment.

> [O]ne often hears people saying that power is that which abstracts, which negates

> the body, represses, suppresses, and so forth. [...] what I find most striking about these new technologies of power introduced since the seventeenth and eighteenth centuries is their concrete and precise character, their grasp of a multiple and differentiated reality. [...] It becomes a matter of obtaining productive service from individuals in their concrete lives. And in consequence, a real and effective 'incorporation' of power was necessary, in the sense that power had to be able to gain access to the bodies of individuals, to their acts, attitudes and modes of everyday behaviour. Hence the significance of methods like school discipline, which succeeded in making children's bodies the object of highly complex systems of manipulation and conditioning. But at the same time, these new techniques of power needed to grapple with the phenomena of population, in short to undertake the administration, control and direction of the accumulation of men[.]
>
> — Foucault 1980, 124-125

Dispositifs function to materialise the *reality* of a power structure. Foucault describes this as the operation of "biopower", a form of power with a specific "technology" for managing populations at large by specifically focusing on disciplining human behaviour. Biopower operates through procedures, technologies, and routines that make life measurable and regulatable, ranging from demographic statistics and health campaigns to public education infrastructures, enabling power to gain knowledge on its subjects, to *gain access to bodies*[2]. Biopolitical formation of dispositifs operationalises the knowledge over bodies in order to produce subjectivities suited to sustaining its specific socioeconomic order. In Foucault's account, this mode of power that embeds itself directly into the conditions of life is the essence of what he names "disciplinary societies".

Institutions such as schools, hospitals, factories, and prisons as Deleuze (1992) argues, "moulded" individuals through enclosure, surveillance, and routine, producing docile subjects whose bodies and conduct could be optimised (see Foucault 1995). However, in "Postscript on the Societies of Control" (1992), he identifies an emerging transformation: from discipline to control. Whereas disciplinary regimes moulded individuals within enclosed institutions, control societies operate through continuous processes that "modulate" individuals by extracting and acting on data traces at the level of what Deleuze calls the "dividual"; subjects fragmented into actionable data particles rather than addressed as unified individuals. Discipline segmented bodies in space and time; control turns their characteristic attributes and behavioural patterns into a subject of analysis across digital networks, data flows, and feedback loops, shaping subjectivities through ubiquitous computational processes rather than architectural confinement. Deleuze's text offers a powerful lens for analysing the late turn of capitalism that Foucault had already begun to trace in his account of neoliberal governmentality (see Foucault 2008). In this formulation, the economy ceases to be one domain among others with its own rationality; it comes instead to encompass the entirety of human action, insofar as all behaviour can be recast as the allocation of scarce resources toward competing ends. What matters is no longer the reconstruction of a mechanical logic, but the analysis of conduct itself as governed by a specific economic rationality (Lemke 2001, 197). The machinery of control, with its computational infrastructure, articulates a future in which such rationalities are operationalised through statistical inference, acting directly on bodies:

> Types of machines are easily matched with each type of society—not that machines

[2] In critical theory, the concept of *body* is usually, though not exclusively, taken to mean the human body, but also understood as a surface of social inscription, always situated in its social context. While philosophy, since Descartes, often mistrusted the body as a source of impulses, Spinoza insisted on asking what a body can do. A major shift came with Merleau-Ponty's phenomenology, which foregrounded embodied perception, and with feminist theory, beginning with Beauvoir, which exposed the neglect of sexual difference. Later thinkers such as Butler challenged the distinction between natural and cultural bodies, and Haraway reconceived the body as cyborg, blurred with animals and machines. Politically, feminism introduced the notion of "body politics," while cultural studies analysed the body as a site of media representation and social anxiety. Foucault's concepts of discipline and biopower remain central, highlighting how bodies are inscribed and governed within regimes of power (see Buchanan 2018, 98–99).

> are determining, but because they express those social forms capable of generating them and using them. [...] capitalism is no longer involved in production [...]
>
> [M]arketing has become the center or the "soul" of the corporation. We are taught that corporations have a soul, which is the most terrifying news in the world. The operation of markets is now the instrument of social control and forms the impudent breed of our masters. Control is short-term and of rapid rates of turnover, but also continuous and without limit, while discipline was of long duration, infinite and discontinuous. Man is no longer man enclosed, but man in debt.
>
> — Deleuze 1992, 5

Deleuze radicalises Foucault's insight by linking neoliberal rationality directly to the machinic infrastructures of late capitalism. The economy does not simply subsume all human action under its logic, but does so through the operational codes of marketing, circulation, and debt, which are engraved in the mechanism of control. Unlike the fixed enclosures of disciplinary institutions, control is continuous, adaptable, and dispersed across networks, markets, and micropolitical dimensions of everyday life. Deleuze's fragmentary diagnosis thus leaves us with an uncanny resonance with today's sociotechnological formations. His concept of control societies has since been expanded in multiple directions: as the organising substance of "Empire" (Hardt 1998); as a framework for analysing digital platforms and sociotechnological imaginaries (Galloway 2001; Raunig 2016; Rouvroy 2012); through studies of surveillance and "dataveillance" (Cheney-Lippold 2017; Haggerty and Ericson 2000; Krasmann 2017); and more recently, through investigations into the institutional roles of emerging technologies and their effects on agency (Amoore et al. 2024; MacKenzie and Porter 2021).

As Michael Hardt (1998, 139) observes, "Deleuze says remarkably little about the institutional architecture of control societies". Some critics even question whether there is a substantive transition from discipline to control at all (see e.g. Kelly 2015), and one might equally doubt whether Deleuze's sketch is an adequate starting point for analysing the social impact of genAI models. My research proceeds, however, from the conviction that Deleuze's account of control societies can be read as a *matrix*[3], not in the sense of a fixed architecture but as the representation of a transformation. What matters is less where each input is mapped than the direction and structure of change itself. In this sense, the Postscript charts the eigenvectors[4] of capitalism: the invariant tendencies along which social investment is regulated. This framing provides a valuable starting point for situating the social role of genAI models. Despite the rich continuation of the literature, scholarship on control societies has often fallen short at critical junctures. It either

i. fails to develop a theory of resistance adequate to the constellation of dispositifs, leaving Deleuze's brief gestures towards lines of flight[5] largely undertheorised;

ii. avoids engaging with the technical machinery of the dispositifs, whether in analysing those described in the definition of control societies or in considering whether contemporary computational infrastructures may already be surpassing them;[6]

iii. and, whether for temporal reasons or due to particular disciplinary focus, neglects AI systems as a primary subject of analysis.[7]

[3] In linear algebra, a matrix represents a function: every linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ can be expressed as multiplication by a matrix A, such that $T(x) = Ax$. A matrix therefore, encodes the rule of transformation rather than the fate of individual inputs, mapping one structured field into another (Strang 2016, 401–420).

[4] An eigenvector of a matrix A is a nonzero vector $\mathbf{v}$ such that $A\mathbf{v} = \lambda\mathbf{v}$ for some scalar λ. Geometrically, the transformation preserves the direction of $\mathbf{v}$ while scaling it by λ, showing how certain tendencies remain invariant even as magnitudes change (ibid., 288–304).

[5] The Deleuzoguattarian term *line of flight* (ligne de fuite) refer to the formation that diverges from the established status-quo's grip, a path that enables parts of a system to break away, reconfigure, or deterritorialise existing structures of power, meaning, or order; "an infinitesimal possibility of escape" (Fournier 2014). A claim "that social formations are defined not by their internal contradictions, but by what escapes them" (Thornton and Holloway 2018, 14).

[6] This point also opens a further discussion: when, if ever, do we move beyond control societies? Is there a field outside them, or does the very ambiguity of Deleuze's formulation prevent us from clearly discerning their borders?

[7] See Galloway (2004) for a work that addresses the other points above, but was published prior to the current surge of developments in AI. See also Section 1.2 for a fuller discussion of Galloway's contribution.

Furthermore, those partly introduced attempts to extend Deleuze's work reveal inconsistencies that sit uneasily with the broader Deleuzian theory. The *Postscript* may be thin in theorising resistance or outlining concrete paths of divergence from the constellation of control societies. Yet Deleuze's wider project, above all "Capitalism and Schizophrenia" with Félix Guattari (see Deleuze and Guattari 1983, 1987), is anything but devoid of revolutionary thinking; quite the contrary, it is constructed around it. Crucially, Deleuze insists that every established power structure already harbours within itself the elements of a resistance against it (see especially the final chapter of "Anti-Oedipus" 1983, 273–383).

This distinction becomes particularly salient when we turn to contemporary algorithmic dispositifs such as genAI. The immanence of resistance in Deleuzoguattarian theory provides a critical alternative to the dominant perspectives on these technologies: it allows us to diverge away from both the uncanny optimism of techno-solutionism and the pessimistic defeatism that frames them as mere tools of techno-feudalist formations. Consequently, my aim is to investigate how these algorithmic assemblages are not merely instruments of modulation but also sites of unforeseen potential. The central question is whether we can move beyond the sparse account of resistance in the *Postscript* to theorise new forms of divergence, even when control seems as subtle and encircling as Deleuze describes. It is precisely here that I turn to D&G's "Capitalism and Schizophrenia". Their work provides the theoretical apparatus that the *Postscript* lacks, enabling an analysis of genAI simultaneously along the mechanism of control and also as a terrain where resistance can be reconfigured. This leads to the study's guiding concern:

> ***RQ***: How are critique and resistance in today's sociotechnological constellation to be (re-)theorised through D&G's project "Capitalism and Schizophrenia", by analysing the emergence of genAI in relation to the institutional framework of control?

The nature and relationship between critique and resistance is hardly new in D&G's work. Yet the central concern of this study starts from a point that can be best described via the question raised by Antoinette Rouvroy (2012): is critique still possible after the "computational turn"? Rouvroy describes a regime in which decision-making no longer depends on politics, law, or social norms, but on data-driven inference, a rejection of modern rationality (see ibid., 2–14; see also Rouvroy 2020), constituting the basis of her notion of "algorithmic governmentality". In such a setting, critique risks being bypassed by predictive infrastructures that act before subjects can intervene. It is precisely here that Iain Mackenzie (2018) offers a different perspective, reframing critique not as an obsolete practice but as a necessary and adaptable one within these new sociotechnological formations:

> Critique as a practice of stepping beyond the limits of possible knowledge, for some, came to replace the idea that critique should establish the limits of legitimate knowledge. [...] we take it that the status of critique in control societies can be positively reframed and that it is necessary to do so if we are to ward off the dangers of a conservative embrace of that which simply concerns us most, or a dogmatic position of commitment in the name of a subject of truth [...] we claim that critique has a history. Not just that it must mobilise historical material to ward off a-historical tendencies, in the manner of historical materialism or genealogy for example, but

> that the very idea and practice of critique must adapt as social formations evolve and change.
>
> — MacKenzie and Porter 2021, 17

Rather than abandoning critique, MacKenzie and Porter (ibid.) emphasise its necessity within contemporary sociotechnological formations. Critique enables more than just stepping beyond current epistemic borders; it functions as the essential force that prevents collapse into indifference, sedimentation, or a-historical regression. In this sense, critique is not merely relevant to resistance but constitutes its very precursor and substance (see I. Mackenzie's earlier work "Resistance and the Politics of Truth" 2018). As much as power/knowledge, resistance/critique is fused and one; they form an inseparable dyad. Critique must therefore adapt to new emergences and develop ways to *step beyond* established configurations. My claim extends this trajectory: not only does critique remain possible in the advent of genAI models, but these meaning-making entities can themselves become instruments of divergence, sources of novelty within processes of subjectivation and human–machine interaction.

1.2 *Neoplatonic Latency: Current Debates and Literature Review*

Contemporary debates surrounding genAI are already abundant. Yet the critique of algorithmic power has a longer genealogy than the recent enthusiasm over generative systems. The rise of predictive analytics inaugurated what some call the "datalogical turn" (Clough and Gregory 2015), a mode of governance grounded in the continuous capture and operationalisation of data traces, while at the institutional level decision-making was analysed under the rubric of "algorithmic governmentality" (Rouvroy 2007). Scholars have shown how these infrastructures align with neoliberal rationalities, translating social life into market norms (Demir 2019), and how they feed into surveillance capitalism (Zuboff 2019). Parallel work has examined the ethical stakes of algorithmic deployment, focusing on fairness, bias, and discrimination in both design and application (Kordzadeh and Ghasemaghaei 2022). Attempts to adapt Deleuze's notion of control to these developments have taken divergent paths. Some extend it to big-data modulation and predictive environments (Brusseau 2020), while others question its adequacy for present formations (Hui 2015). Earlier accounts of "dividualisation" (Cheney-Lippold 2011; Van Otterlo 2013) remain important, but they predate today's generative infrastructures and therefore cannot address how meaning production itself becomes a site of governance. While critiques of algorithmic governance are extensive, they often centre on surveillance, datalogical rationality, or ethical bias without engaging the representational novelty of genAI. The capability of these models, especially Large Language Models (LLMs), to perform meaning-making (Dishon 2024; Gretzky 2024; Mishra and Heath 2024) has provoked renewed inquiry. By traversing vast data foundations to generate outputs that appear plausible within a learned distribution, they reposition representation itself as a site of governance, even as their operations remain opaque in causality and justification. Although the theoretical literature on genAI is still thinner than that on earlier algorithmic systems, several contributions stand out as particularly relevant for this study.

In the genAI related literature, Louise Amoore et al. (2024) shift attention from surface-level failures to the deeper representational structures of AI. While Emily M. Bender, Timnit Gebru et al. (2021) warned that LLMs risk "parroting" entrenched arguments in their training data, a risk amplified when models are trained on their own outputs, Amoore et al. argue that the real concern is how models generate a *political* "world model"[8]. Rather than simply repeating arguments, these systems subtly orient data interpretation toward ideological directions that are difficult to foresee. Although contemporary genAI models remain far from constructing a genuine world model, Amoore et al. express concern that the central representations these systems derive from their training data increasingly begin to fill the gaps opened during generation (in the case of LLMs, for instance, gaps of meaning), while the model compresses and interprets data into lower-dimensional representational spaces ("latent spaces"). Yet this focus, while introducing important questions about the nature of human–machine communication, leaves aside how and why these abstractions of the world derived through data come to take monolithic form (see Section 3.3). It also remains unclear how these representations differ from human cognition, which likewise depends on schematic reductions of experience. Despite these limits, Amoore et al.'s critique links world-modelling to debates on "machinic Neoplatonism" (Eloff 2021; McQuillan 2018), the idea that reality is best perceived mathematically, waiting to be extracted from data. Building on this, Aragorn Eloff (2021) refers to the term "Algocene"; a new epoch shaped by algorithmic environments, atemporal, high-speed networks that fragment subjectivity into dividual aggregates. The novelty of Eloff's formulation lies in showing how these environments shape human cognition in ways that resemble the operations of deep learning systems, where contemporary forms of communication take on the structure of adversarial pattern-recognition and recursive error correction. It is this distinctive configuration that Eloff (ibid., 179) names the "algoplastic" stratum. While drawing on Deleuze's *Postscript*, he is concerned less with specifically deliberative behaviours of genAI models than with the wider transformations of subjectivity that arise when humans interact with increasingly pervasive algorithmic systems in a continuous bilateral exchange that reorganises political discourse. His account also highlights the cognitive vulnerabilities humans suffer from in encounters with *other* meaning-making entities, offering a productive angle for examining how agency emerges within the Algocene. The problematic of agency is taken up further in accounts of the sociotechnological imaginary of AI, whereas Gideon Dishon (2024) and Paul Prinsloo (2017) emphasise how interactions with genAI unsettle clear boundaries of action and intention. Dishon critiques the dominant "Frankenstein" imaginary, which frames AI as an anthropomorphic, external threat, and instead proposes Franz Kafka's "The Trial" (1988) as a more apt metaphor. In this view, agency becomes entangled: humans and genAI systems are drawn into a mutual, ongoing attempt to interpret and respond to each other without ever fully grasping the other's logic. Relations are recursive in a specifically interpretive sense, as both sides continually generate and infer meaning in ways that reshape the encounter itself. Control, rather than appearing as a struggle over the inscription of subjectivity, shifts toward patterned constraint, where genAI expands the set of choices available to users even as it narrows the forms of meaning that can emerge within those choices, producing an extended and indeterminate process of negotiation. Dishon's intervention offers

[8] The notion of a "world model" comes from current AI research, where the next frontier is framed as the development of models capable of learning richer, actionable, and more versatile representations of their environment (LeCun 2022a). Such an advancement is expected to enable systems to generalise to previously unseen problems (outside of the model's training data or previous experience) and produce context-sensitive solutions in ways that appear intuitive.

a distinctive way of understanding how genAI reorganises the conditions under which meaning is produced, and clarifies why familiar dichotomies of agency, relations, and control no longer hold in the Kafkaesque landscape of contemporary human–machine interaction.

A common thread running through these debates is that both meaning and agency are shaped through infrastructural arrangements that exceed individual interaction. This shift becomes explicit in the analysis developed by MacKenzie and Porter, who define emergent genAI infrastructures as computational "totalising institutions". Their account highlights how AI models sequence dividual traces across domains, with scores, profiles, and categories continuously inscribed, routed, and repurposed so that conduct is modulated over time (see MacKenzie and Porter 2021, 23–24). While these reflections echo earlier analyses of algorithmic governance, their distinctive contribution lies in linking critique and resistance through the concept of "counter-sequencing":

> While these algorithmic functions are now well known, and can be critiqued at the level of the potentially infinite process of signification they constrain, they can also be critiqued through a process we would call counter-sequencing. Counter-sequencing is the activity of reordering the power diagram of the totalizing institution in ways that destabilize its functioning. That said, it would be unwise to assume in advance that counter-sequencing must result in some kind of 'positive' ethico-political outcome. The aim, instead, is to understand the critical potential of counter-sequencing first and then to engage in, what Williams calls, the revaluation of that critique with more 'local', that is 'pragmatic', concerns at the forefront of such revaluations.
>
> — ibid., 23–24

Mackenzie's proposal is significant for shifting the focus from critique as diagnosis to critique as intervention. Yet it remains underspecified; counter-sequencing is described more as a gesture than as a concrete practice, leaving open questions about its operational form and its place in processes of subjectivation.

By contrast, Federico Montanari (2025) addresses these gaps more directly by engaging the technical mechanisms of contemporary architectures. He foregrounds procedures such as *dimensionality reduction* (similar to Amoore et al.'s (2024) reference to latent spaces) and the *transformer architecture* underpinning genAI, analysing how its technical operations produce cultural resonance. Montanari argues that transformers exemplify the interplay between metaphor and function: as specialised NNs that simulate certain brain structures, they excel at processing sequential data through their *attention mechanism*, an innovation that enables selective focus on relevant parts of input sequences to discern complex relationships and dependencies within data (Montanari 2025, 206). This functionally specific yet mythically resonant architecture, he suggests, reveals how technical metaphors solidify both the utility and mystique of AI systems. Here, resistance is framed not as wholesale rejection but as a possibility internal to machinic processes themselves. Montanari (see ibid., 208-210) notes that the agency issue might be more complex than it seems, and emphasises the necessity for research into the inner working of the genAI models in order not to leave the understanding of their capabilities and future development to tech giants. His account, however, remains largely programmatic, leaning towards ethical risk and cultural

semantics rather than concrete pathways for architectural reconfiguration.

It is here that Pierre Beckmann et al. (2023) advance the discussion with a decisive technical intervention. They propose "computational phenomenology", a framework that rejects the "neuro-representationalist" assumption that AI systems interact with the outer world only through a monolithic representation they build. Instead, they argue that genAI should be understood as a processual style of sense-formation: it generates meaning not by referencing a stored map of reality, but through dynamic, context-sensitive activations within its neural network. This perspective highlights the potential of repurposing and reconfiguring pretrained networks by emphasising or capitalising on, for example, specific layers of meaning formation. The "DeepDream" experiment can serve as a metaphorical example: a model initially intended as a classification architecture was turned into a generative model (ibid., 416), ultimately producing novel and dream-like images. Read this way, architectural variation itself becomes a site for configuring models differently, or for actions such as counter-sequencing that displace the centripetal pull of standardised "world models" and open alternative structures. Taken together, these accounts indicate both the richness and the limits of contemporary theorisation. While MacKenzie and Porter introduce counter-sequencing as a critical method, Montanari and Beckmann et al. point toward technical and conceptual openings for resistance within genAI. Yet their analyses remain partial. In what follows, I extend these debates by combining political-theoretical insights with a closer examination of generative infrastructures, in order to develop a framework for critique and resistance adequate to today's algorithmic dispositifs.

1.3 *Methodological Approach*

Cere-nominal, a concept from William S. Burroughs' (2012) "The Soft Machine".

The methodology follows from the gaps identified in existing scholarship, adopting an analytical, interdisciplinary, critical, and technical orientation. Returning to the academic approach suggested by Alexander R. Galloway (2004):

> For Empire, we must descend instead into the distributed networks, the programming languages, the computer protocols, and other digital technologies that have transformed twenty-first-century production into a vital mass of immaterial flows and instantaneous transactions. Indeed, we must read the never-ending stream of computer code as we read any text (the former having yet to achieve recognition as a "natural language"), decoding its structure of control as we would a film or novel.
> — Galloway 2001, 82

Although writing at a much earlier stage, Galloway (see Galloway 2001, 2004) was among the first in critical theory to examine the technical machinery of the internet in order to show how control persists even after decentralisation. While his work plays a contextual role in this study, his methodological orientation also provides an important inspiration for how to approach the present research.

In analysing genAI within the framework of control societies, the methodology proceeds on two fronts. First, it traces the historical development of AI, with particular attention to the breakthroughs that enabled contemporary models, most notably the transformer architecture. Second, it situates these systems institutionally, examining their embedding within structures of power and knowledge.

genAI is therefore analysed not merely as a technical artefact but as part of a wider constellation of dispositifs that reorganise governance and subjectivation.

Accordingly, the approach combines:

i. a genealogical reading of power that situates genAI within the shift from disciplinary to control societies,

ii. a close technical exegesis of model architectures, with attention to how specific mechanisms condition the politics of mediation,

iii. an examination of the institutional embedding of contemporary computational infrastructures,

iv. and a reflection on how resistance and critique may be theorised through D&G's "Capitalism and Schizophrenia".

Ultimately, I argue that while the generative capabilities challenge the pillars of the control society concept, they find particularly insightful correspondences in other literature of D&G.

1.4 The Cartogram

The thesis adopts a structure that resists straightforward sequential progression. Each chapter forms a distinct yet connected zone of inquiry, linked to others through cross-references that allow arguments to circulate rather than accumulate in linear order. To prevent ambiguity within this architecture, chapters incorporate brief orientations and concluding summaries, enabling them to function as semi-autonomous units. The Cartogram that follows outlines this arrangement, mapping how the project's core concerns, genAI as an institutional dispositif and the possibilities of critique and resistance, are distributed and interlinked within this constellation.

Chapter 2 reconstructs the conceptual shift from disciplinary to control societies, situating genAI within this transformation. It begins with Foucault's genealogy of subjectivity, showing how disciplinary institutions like prisons, schools, and factories moulded docile subjects through enclosure, surveillance, and routine. Deleuze's *Postscript* marks a rupture: the emergence of control societies that act through continuous modulation, dividuation, and algorithmic circulation rather than architectural confinement. This provides the theoretical ground for analysing genAI not simply as a technical tool, but as a dispositif that reorganises governance and subjectivation by acting on data traces instead of enclosed bodies. The chapter revisits key concepts such as docility, modulation, and dividualisation to clarify how subject formation evolves under institutional and computational infrastructures.

The second part develops the problem of critique and resistance. Deleuze's Postscript leaves resistance underspecified, offering only fragmentary hints in his reference to "program". The chapter, therefore, turns to later theorists such as Hardt, Negri, Galloway, and MacKenzie & Porter, who expand Deleuze's formulation while also introducing a notion of resistance. By bringing these debates

into contact with genAI, the analysis shows that critique cannot be external to control but must operate within its infrastructures. Resistance is theorised less as external negation than as the activation of lines of flight and micropolitical apertures already present in algorithmic dispositifs. This sets the stage for subsequent chapters, which deepen the analysis through technical, institutional, and machinic perspectives.

Chapter 3 shifts the focus to the technical and historical development of AI, tracing the path from Symbolic Artificial Intelligence (symAI) to contemporary Deep Learning (DL) and generative models. The chapter first outlines the classical paradigm of symAI, describing how reasoning was formalised in logical rules and arborescent systems, as well as why these approaches ultimately fell short in handling complexity, ambiguity, and contextuality. It then turns to the rise of connectionist models, beginning with early NNs and their limitations, before examining the breakthroughs in DNNs and Self-Supervised Learning (SSL) that underpin today's architectures.

The analysis emphasises how these technical shifts are not only engineering milestones but also epistemological transformations. With the advent of *transformer architectures* and *attention mechanisms*, genAI systems acquired the capacity to generate rather than merely predict, reorganising the role of AI from a tool of decision support to an infrastructure of meaning-making. The chapter situates these developments within broader debates, showing how probabilistic inference, tokenisation, and optimisation procedures reconfigure knowledge production and subjectivity. The analysis dives deeper into the machinery that made the contemporary genAI models exceptional, like *transformer architecture*, how under-/overfitting are handled, how *gradient descent* and *backpropagation* work. These functionalities at times are analysed through D&G's theory. This historical and technical mapping provides further foundation to reflect on current debates and conceptualisation (and/or verification) of specific (re-)configuration methods for genAI models.

Chapter 4 shifts from architectural analysis toward the problem of agency, examining how contemporary genAI systems reshape the conditions under which meaning, interpretation, and critique become possible. It opens with concerns over the representational limits of LLMs, drawing on Bender, Gebru et al. and Amoore et al. to show how statistical reconstruction and algorithmic inference risk naturalising political or ideological tendencies within predictive infrastructures. The discussion then turns to Eloff's notion of the algoplastic stratum, in which humans and genAI models interact within a shared interpretive space characterised by opacity, recursive sense-making, and shifting boundaries of agency. This provides the basis for incorporating Dishon's argument that genAI introduces a Kafkaesque dynamic, where action and intention blur through ongoing negotiation between human and machine.

From there, the chapter explores alternatives to representational readings of AI, engaging Beckmann et al.'s "computational phenomenology" as a framework in which meaning arises not from fixed internal maps but from layered activations that unfold contextually within NNs. Montanari's contribution adds up

to this trajectory by highlighting transformers' capacity for long-range conceptual relations and speculating on futures in which genAI participates more actively in socio-political narration. Taken together, these perspectives frame human–machine interaction as a hybrid and evolving formation, where meaning is continuously assembled across technical and social strata. This conceptualisation opens space for resisting convergent rationalities by intervening in the layered processes through which genAI systems generate patterns, narratives, and modes of subjectivation.

Chapter 5 brings together the technical, institutional, and theoretical strands developed in the previous chapters in resemblance of conjunctive synthesis. The chapter argues that generative architectures bind together heterogeneous forces: data, algorithms, user inputs, institutional logics, and cultural imaginaries; into operational wholes, while still preserving indeterminacies that can be activated for divergence. The chapter begins with a microphysics of resistance, showing how model behaviour such as hallucination, chain-of-thought expansion, coherence production, and alignment not only stabilise meaning but also instantiate deviations and misfires that resist full capture. These deviations are read as micropolitical opportunities for activating creativity, divergent thought, and even intention within the model.

The chapter then offers a series of interventions that aim to prevent generative architectures from becoming sedimentary. Techniques such as feature amplification, artificial curiosity, and the introduction of deliberate perturbations are framed as methods for sustaining a non-conforming tendency inside the model. The chapter uses the concepts from "Capitalism and Schizophrenia" (Deleuze and Guattari 1983, 1987), like schizoanalysis and nomadology, to articulate how the potential for divergence already pointed out in the technical analysis of generative systems can be activated. The aim is neither romantic refusal nor technophilic celebration but a pragmatic activation of alternative sense-configurations within the architecture itself. The final part of the chapter develops the concept of counter-sequencing as a way of reorganising the power diagrams of genAI. Counter-sequencing treats hallucination, divergence, and improvisational generation as resources for reshaping the structures of human–machine interaction. Chapter 5, therefore, provides a conceptual and practical framework for navigating critique and resistance not as a refusal of generative infrastructures but as an immanent practice within them.

Precisely by pursuing these prospects, I argue that the specificity of generative architectures is overlooked by the dominant perspectives, which either reduce them to traps of subjectivation or celebrate them through techno-solutionism. My work argues that both positions overlook the specificity of generative architectures. GenAI not only reterritorialises meaning by producing coherence from fragmented inputs; it also contains, within those same cohering processes, the potential for divergence. As computational dispositifs become more sophisticated, so too do the micropolitical possibilities for counter-sequencing, divergence, and intervention.

The significance of this project lies in addressing two persistent gaps. First, Deleuze's sketch of control societies leaves the question of resistance under-

theorised; my analysis extends this by grounding lines of flight in the concrete operations of genAI. Second, much scholarship on algorithmic power remains at the level of metaphor or ethics, without engaging the architectures themselves. By situating critique within the technical and institutional logics of transformers and related models, the study develops a framework for understanding generative infrastructures not merely as instruments of capture, but as terrains where new subjectivities and forms of political action can emerge.

2

Subjectivity under Control: Critique and Resistance in Post-Disciplinary Societies

> A systematic rejection of subjectivity in the name of a mythical scientific objectivity continues to reign in the University. In the heyday of structuralism the subject was methodically excluded from its own multiple and heterogeneous material of expression. It is time to re-examine machinic productions of images, signs of artificial intelligence, etc., as new materials of subjectivity.
>
> Félix Guattari 1995a, 133

Gilles Deleuze's (1992) short and speculative essay, "Postscript on the Societies of Control", introduced a fragmentary but generative diagnosis of contemporary power structures. It sketches a transition from the closed environments of institutions like school, factory, prison that played the pivotal role in shaping human subjectivity into a more diffuse machinery, increasingly reinforced by digital technologies and governed via the functionalities made available by computational advancements. In control societies, these new affordances form an assemblage that distributes the disciplinary operation across the social field without any further institutional mediation (see Hardt 1998, 139), enabling a more fluid, flexible, and continuous operation of power. Control is thus encountered as a micropolitical force acting directly upon everyday life, and is rather characterised with a pre-emptive role.

The formulation of control has already inspired critical readings of emerging computational cultures, internet infrastructures, and surveillance capitalism. Early developments in Artificial Neural Networks (NNs) and Deep Artificial Neural Networks (DNNs) have already played a prominent role by enabling increasingly capable systems, built upon a long history of Natural Language Processing (NLP), and supporting applications such as search engines, social media platforms, recommendation systems, and automated filtering. Within this trajectory, Artificial Intelligence (AI) quickly became a central object in theorising control societies; yet, we now stand at a threshold beyond the early imaginaries of *cyberspace* or the *virtual*: the contemporary AI landscape is dominated by Generative Artificial Intelligence (genAI) systems, particularly Large Language

Models (LLMs), which no longer merely transmit or classify information; they increasingly participate in the production of meaning itself (Dishon 2024; Kazakov 2025). These models generalise across domains, transfer knowledge between tasks, and adapt to unforeseen situations rather than remaining bound to narrow, predefined functions (Xu 2024). Kazakov (2025) characterises this development as a mode of scalar Darwinism, defined by relentless quantitative expansion rather than qualitative transformation. LLMs and other genAI systems advance primarily by scaling; more parameters, larger datasets, and increasing computational resources, without fundamental architectural innovation. This trajectory reinforces existing capitalist logics, treating data as a resource to be extracted and leveraged; competitive advantage derives from scale rather than novelty. Just as neoliberal governmentality construed *the market* as a quasi-metaphysical plane that produces the optimal outcomes without direct intervention (Foucault 2008, 131), contemporary AI discourse often assumes that scaling models and data will automatically yield the solutions humanity is said to need.[1]

Whether this acceleration will eventually enable more sophisticated forms of AI, including Artificial General Intelligence (AGI)[2], remains contested. What is clear is that these systems continue to expand in capability, generality, and reach. GenAI models, especially LLMs, now function as computational agents that operate within and across domains, mediating how information is organised, circulated, and apprehended. They do not simply support existing knowledge practices; they increasingly generate outputs that are taken as meaningful, authoritative, and actionable (Dishon 2024; Montanari 2025). Their interpretive operations shape what becomes visible or legible, filter which forms of knowledge can travel, and condition how subjects encounter and interpret information. Processes of subjectivation, therefore, unfold within a landscape where computational models actively participate in producing the categories, associations, and interpretive cues through which reality is navigated. As genAI systems occupy roles once associated with expert judgement, they begin to function as distributed institutional actors. This development intensifies the micropolitical dynamics identified in control societies, raising the question of how critique and resistance might be articulated when meaning is co-produced by systems whose authority derives from scale and statistical inference. The following analysis situates genAI within the historical transition from disciplinary institutions to control, in order to examine how these models shape institutional logics and the production of subjectivity.

The chapter proceeds in four steps:

1. It revisits the genealogy of subjectivation, from classical accounts to post-structuralism.
2. It updates Deleuze's account by examining modulation, dividuation, and the role of computational infrastructures in extending discipline into continuous control.
3. It considers how critique and resistance may be conceptualised under these conditions, engaging some of the relevant reflections on Deleuze's (1992) formulation of control societies.
4. It then turns to contemporary debate around genAI models and examines how

[1] Not always this explicit, but the technosolutionist propagation is particularly strong in the frontlines of tech giants:

— Elon [@elonmusk] Musk 2025

[2] AGI is a hypothetical intelligence of a machine capable of understanding, learning, and performing any intellectual task that a human being can do. It generalises across domains, transfers knowledge between tasks, and adapts to new, unforeseen situations, rather than being specialised for narrow tasks (Xu 2024).

they may function as institutional dispositifs that govern meaning and what might be their relevance to the already established arguments on resistance and critique (or the lack thereof), preparing the scene for an in depth analysis in the following chapter.

2.1 *The Genealogy of the Docile Bodies*

As Foucault (1978) defined the trajectory of politics as approaching the desire to infuse the "life itself", he framed the pursuit of an effective methodology of subjectification[3] axis around which "biopower" and "biopolitics" were organised.[4] The problem of subjectivity; its emergence, formation, and nature, however, has long occupied Western philosophy, morphing with shifts in different branches of philosophy like epistemology and metaphysics. Its philosophical genealogy stretches back to ancient concerns with soul and selfhood (see e.g. Aristotle's "De anima (On the Soul)" 1986), but its modern formulation takes decisive shape with René Descartes's (see 2008) "cogito", which installs the thinking subject as the indubitable ground of knowledge. From there, Immanuel Kant's (see 2009) *Copernican Revolution* redirects philosophical inquiry by elevating the mind from its passive definition to an entity that actively structures our experience of the world, putting it at the centre of our perception of the world. Subjectivity is already playing a prominent role in Kant's theory as we perceive the world as it appears to us (in phenomena) and not as it is (in noumena), however, it is Georg Wilhelm Friedrich Hegel (see 2019) who emphasises the social substance of subjectivation (which can be read as a part of his *Keplerian Revolution* in philosophy). Hegel shows that the subject comes into being only through social struggle and dependence, binding subject formation to power from the outset. It is therefore no coincidence that Karl Marx, especially in his earlier works, emphasises the role of labour as a medium of human subjectivity (e.g. in Marx 1988). In these writings, Marx critiques idealist accounts of subjectivity, arguing that subjectivity arises from "sensuous human activity" that is always shaped through socio-historical development and grounded in material conditions rather than in an abstract interiority (see Wang 2023, 3). This can be read as standing in relation to his broader attempt to turn Hegel's dialectic upside down and place it on a materialist foundation, grounding the becoming of the subject in concrete practices and the historical life-processes of species-being[5] (see Wang 2023, 3).

Thus, even before twentieth-century critiques, the entanglement of subjectivity and power is already visible. Following Marx, Georg Lukács deepens the problem of subjectivity by arguing, through his reading of early Hegel, that labour mediates the unity of subject and object, creating a "second nature" in which the subject forms itself through socio-historical activity (see Lukács 1976). In both his early writings and his later works, Lukács treats labour as the practical medium through which humans transcend mere instinctual life, realise their purposes in the world, and achieve relative freedom within material constraints (see Wang 2023, 4). By grounding subject formation in concrete social practice, Lukács marks the final major account centred on labour before later twentieth-century theories would fundamentally challenge the very primacy of such a humanistic, agent-centred subject. Works of Ferdinand de Saussure (see 2011), Claude Lévi-Strauss (see 1963), and Louis Althusser (see 1977) have shifted attention from in-

[3] The terms subjectivation (although the British English form of the word subjectivisation would be more fitting, critical theorists, especially French scholars, tend to use this form more often) and subjectification are often used interchangeably. However, this study refers to subjectivation in terms of Hegelian becoming, that is, becoming conscious as an internal process; and subjectification as an external process that refers to the formation of subjectivity through a machinery, as, for example, how Foucault mentions it, to emphasise a specific nuance. Still, considering how critical theory, especially post-structuralist theory, refers to subject, subjectivity, and individuality, there is a large intersection between the two. For a more in-depth reading about the difference between the two terms, refer to Chapters 3 and 5 in Wille et al.'s (2015) "Spaces and Identities in Border Regions".

[4] Although Foucault introduces these terms early in his theory (see e.g. 1995), for a more concrete definition in relation to neoliberal governmentality, refer to "The Birth of Biopolitics" (2008).

[5] Alternatively species-essence; Ger. *Gattungswesen*. The concept Marx (see 1988) uses to refer to the essential social and creative nature of human beings, which is realised through free, conscious activity and social relations.

terior experience to the impersonal systems of language, myth, and ideology that precede and produce subjectivity. The subject, in this specific current, becomes an effect of signifying structures; it is interpolated by ideological apparatuses and made legible within symbolic orders. Post-structuralist thought introduced a specific kind of radicalisation of it: Roland Barthes (see 1977), Jacques Derrida (see 2016), and Julia Kristeva et al. (see 1980) foreground the instability, iterability, and difference at the heart of these structures themselves. In post-structuralism, the autonomous subject of modernity dissolves into the relational field of discourse and social practice. What appears as subjectivity is only an instance, a provisional effect emerging from the entangled formations of language, power, and social formations. There is no pre-given subject behind the *text*; the subject exists only as a position; produced, fractured, and dynamic.

Where post-structuralism dissolves the subject into discourse, Foucault resituates its production within the material operations of power, showing how institutions and practices fabricate subjects through discipline. In his account on "the careful fabrication of subjectivities" (Foucault 1995, 215), disciplinary institutions operate as enclosed environments that shape individuals by acting directly on their bodies and conduct. Schools, prisons, factories, and families regulate behaviour by organising routines, structuring space, assigning tasks, and training bodies[6] to perform specific functions. The mechanism of discipline relies on continuous surveillance. The "panoptic" apparatus induces a "state of conscious and permanent visibility" (Foucault 1995, 202–203), prompting individuals to monitor and correct themselves as they internalise the institutional gaze. Unlike sovereign power, which relied on spectacular punishment and the prerogative "to rule on death rather than to administer life" (1992, 3), disciplinary power works through the everyday regulation of bodies and habits. It functions as a "political technology of the body" (Foucault 1995, 26): diffuse rather than centralised, operating through innumerable practices, spaces, and procedures. Since it is enacted by institutions and reproduced by the subjects themselves, discipline exhibits what Foucault calls a "microphysical manifestation of power" (ibid., 26–27). Power therefore circulates throughout the social field rather than emanating from a single source, constituting individuals by making them visible, knowable, and governable.

As the emergence of subjectivity becomes increasingly associated with the exercise of power, modern forms of governance come into focus. They do not merely constrain individuals from the outside but shape what can be known, felt, and done by organising the very conditions of selfhood. Once genAI models participate in producing meaning and participating in the (re-)production of knowledge, the question of how they relate to subjectivation becomes unavoidable. This is particularly relevant given the association Foucault builds between the process of subjectification and subjugation:

> [A]ll these present struggles revolve around the question: Who are we? They are a refusal of these abstractions, of economic and ideological state violence, which ignore who we are individually, and also a refusal of a scientific or administrative inquisition which determines who one is. [...] This form of power applies itself to immediate everyday life which categorizes the individual, marks him by his own individuality, attaches him to his own identity, imposes a law of truth on him which he must recognize and which others have to recognize in him. It is a form of power which makes individuals subjects. There are two meanings of the word "subject":

[6] In critical theory, the concept of *body* is usually, though not exclusively, taken to mean the human body, but also seen as a surface of social inscription, often thought in its social context. While philosophy, since Descartes, often mistrusted the body as a source of impulses, Spinoza insisted on asking what a body can do. A major shift came with Merleau-Ponty's phenomenology, which foregrounded embodied perception, and with feminist theory, beginning with Beauvoir, which exposed the neglect of sexual difference. Later thinkers such as Butler challenged the distinction between natural and cultural bodies, and Haraway reconceived the body as cyborg, blurred with animals and machines. Politically, feminism introduced the notion of "body politics," while cultural studies analysed the body as a site of media representation and social anxiety. Foucault's concepts of discipline and biopower remain central, highlighting how bodies are inscribed and governed within regimes of power (see Buchanan 2018, 98-99).

> subject to someone else by control and dependence; and tied to his own identity by a conscience or self-knowledge. Both meanings suggest a form of power which subjugates and makes subject to.
>
> — Foucault 1982, 781

Subjectivation, therefore, involves a double movement: the subject emerges only through the operations that simultaneously render it governable, and at the same time, governance aims to align emerging forms of subjectivity with its prevailing order of truth. Power is a productive force in this sense; it shapes identities and establishes the standards by which individuals become intelligible to themselves and to others. Production of knowledge is central to the architectural logic of the disciplinary power; as Krasmann notes, it is "not so much about discovering the truth, but rather about producing certain truths" (2017, 11). The knowledge on bodies renders them accessible and open to intervention, and control over the production of knowledge shapes the nature of social production and the truth itself. The disciplinary production of subjects through enclosure and surveillance establishes the conceptual ground for the emergence of a different configuration of power. As the routines, procedures, and techniques of discipline diffuse beyond the walls of institutions, they begin to operate more flexibly, no longer enclosing subjects but acting on them through circulating flows of information, assessment, and anticipation, marking a further shift in their operation.

2.2 *The Emergence of Modulative Control*

> You see control can never be a means to any practical end... It can never be a means to anything but more control... Like junk...
>
> Burroughs 1979, 81

Following Foucault's genealogy of power, from sovereign regimes to disciplinary societies, Deleuze introduces a further historical configuration: "the society of control" (Deleuze 1992). Deleuze points to an institutional crisis and charts the replacement of enclosed institutional spaces; schools, factories, prisons, family[7] by diffuse, pervasive mechanisms of flexible forms of control. Deleuze notes that the 20th century marks the transition, the disciplinary institutions were already fading out after WWII (Deleuze 1992, 3). While disciplinary regimes operated through enclosures, segregating individuals into clearly defined spaces associated with specific functions, control societies rely on more fluid mechanisms: instead of physical boundaries, social organisation is achieved by tracking, directing, and modulating movement and behaviour across interconnected and permeable environments (see Brusseau 2020, 3). As Michael Hardt (1998, 139) puts it, "the walls of the institutions are breaking down in such a way that their disciplinary logics do not become ineffective but are rather generalized in fluid forms across the social field" as new forms of biopolitical governance start taking hold (see also I. Mackenzie 2018, 122).

[7] Which Gilles Deleuze & Felix Guattari (D&G) were eager to emphasise its institutional nature of (see "Anti-Oedipus" 1983).

As movement between institutions required subjects to reset themselves and start anew in each, the machinery of discipline operated in parallel but remained com-

partmentalised. Although a common language circulated between these enclosures, their relation was strictly "analogical" (Deleuze 1992, 4). Control, a term Deleuze borrows from William S. Burroughs (1979), signifies both this institutional shift and a fundamental change in the machinery of subjectivation. It marks the rise of a post-disciplinary dispositif powered by computational advancements, in which power is no longer exercised by confining bodies within discrete environments but by continuously modulating informational flows that shape conduct directly:

> Control differs from governmentality in two interconnected ways: first, control is dependent upon electronic technology as its primary mode of delivery. This most obviously means technology that is able to process, store, and transmit huge amounts of information, creating an "informational milieu" that conditions both what is capable of being thought and how it is to be thought. Second, control no longer presupposes an "outside," not even as an included lack or absence because, operating by differing from itself, it is able to treat the whole of life as so many statistical variations of itself.
>
> — Moore 2007, 457

Building on these distinctions, control can be understood as a regime in which the mechanisms of disciplinary governmentality become internalised within technological infrastructures that act upon continuous informational flows. "Modulation" replaces the "moulds" that characterise disciplinary institutions: where moulding imposed form from fixed sites, modulation operates as a flexible and dynamic process that acts across domains rather than through segmented enclosures (see Deleuze 1992, 4). It functions through ongoing feedback and adjustment, shifting from a "form-imposing" to a "self-regulating" mode in the production of subjectivity (Hui 2015, 74).[8]

Under control, the subject who once moved between distinct institutional roles undergoes a new kind of fragmentation. The self may remain socially continuous, yet its qualities and behaviours are isolated, analysed, and acted upon in smaller, detachable components (MacKenzie and Porter 2021, 5). To capture this transformation, Deleuze introduces the figure of the "dividual": the former individual is now treated as divisible and can be decomposed into data particles, micro-traces, and partial qualities circulating across digital and organisational infrastructures (Deleuze 1992). These dividual elements are treated as operational units; detached from the person from whom they originate, they are mobilised for increasingly automated decision-making processes (ibid., 6). John Cheney-Lippold illustrates the practical consequences of this logic in contemporary, highly digitalised environments:

> [M]odulation marks a continuous control over society that speaks to individuals in a sort of coded language, of creating not individuals but endlessly sub-dividable 'dividuals' [...] Dividual fragments flow across seemingly open and frictionless networks and into rigid database fields as part of the subsumption implicit in data mining [...] As a user travels across these networks, algorithms can topologically striate her surfing data, allocating certain web artifacts into particular, algorithmically-defined categories like gender. The fact that user X visits the web site CNN.com might suggest that X could be categorized as male. And additional data could then buttress or resignify how X is categorized. As X visits more sites like CNN.com, X's maleness is statistically reinforced, adding confidence to the measure that X may be male

[8] Control and its modulating form are not necessarily digital; the term refers to an operational logic that becomes dominant with contemporary capitalism. Deleuze illustrates this shift in the transition from factory to corporation, where production is no longer confined to a delimited site but distributed across an abstract field of work with variable remuneration. Likewise, education no longer concludes with the school but extends into lifelong training (Deleuze 1992, 6). While modulation predates digital systems, contemporary technologies intensify and expand this logic, providing the material conditions through which discipline is transformed into control.

— Cheney-Lippold 2011, 168-169

This configuration produces what MacKenzie and Porter describe as a "bundle of elements held together" (MacKenzie and Porter 2021, 6), a formation that gradually replaces the individual as the primary unit of governance. Through databases, ubiquitous computation, and statistical inference, these dividual traces are parsed, recomposed, and acted upon, generating personalised evaluations, outputs, and interventions. The effect is akin to a "self-deforming cast": a continuously adjusting apparatus that modulates the subject in motion (Deleuze 1992, 4). Docility under this new regime is no longer enforced by explicit institutional intentionality operationalised as rigid codes. Instead, control operates by creating spaces that feel open and permissive, as if the individual were free to explore, create, and tangle with possibilities. Yet both their production and their ends are subtly governed by intangible, underlying forces (Hui 2015, 75) acting on a much more personal level. Control converges the previously separate spaces of subjectivation into a single, fluid field: "one no longer leaves an institution behind, and one is never fully done with the spaces that act upon the self" (Deleuze 1992, 6). Contrary to disciplinary institutions, which segmented individuals and populations, control societies separate components of individuality (MacKenzie and Porter 2021, 9).

As Michael Hardt and Antonio Negri (2003) observe, "the passage to the society of control does not in any way mean the end of discipline. In fact, the immanent exercise of discipline [...] is extended even more generally in the society of control" (also in Galloway 2001, 83). Here, immanence refers to how discipline becomes embedded in circulating processes rather than emanating from fixed institutions, so to say, "subjectivities are still produced in the social factory" (Hardt 1998, 149) but intensified and generalised, going beyond (but not abolishing or replacing, rather extending or coexisting with) the institutions. Control unfolds across different modalities and novelties; in protocols, through feedback loops, on algorithmic infrastructures; governance operates through continuous modulation that conditions how subjects perceive, act, and desire. Yet while Foucault never postulated a stage beyond disciplinary societies, Deleuze's *Postscript* also offers only a sparse sketch of what comes after enclosure-based institutionalisation, the form remains vague, its contours merely suggested through keywords like *modulation* or *dividuation*. What Deleuze leaves us with, then, is not a blueprint but a sketch of tendencies whose operative principles demand further analysis. The challenge, as subsequent debates have often emphasised, is to determine how such a configuration might still allow for critique and resistance. If discipline persists under control in more diffuse and continuous forms, any account of power must also attend to its internal lines of tension and the micropolitical openings through which divergence can emerge.

2.3 *A Critique of Critical Lack of Critique*

If, however, the regime of Power/Knowledge[9] is indeed getting more and more encircling and unified through the new capabilities, new technologies of power; is there any way to diverge, counter-act, open new planes for a different kind of subjectivation? We are landing at the very foundations of the critical theory;

[9] As Foucault (see e.g. 1980) likes to refer it to emphasise the the inseparability and interplay of knowledge and power; power is based on knowledge, operates on knowledge, and in turn power also (re)produces knowledge, shapes knowledge.

as an exemplary articulation, Foucault is positioning the relationship between resistance and critique via referring to Immanuel Kant's (1784) "Beantwortung Der Frage: Was Ist Aufklärung?"[10]:

[10] "An Answer to the Question: What is Enlightenment?"

> [I]n relation to Aufklärung, critique for Kant will be that which says to knowledge: Do you really know how far you can know? Reason as much as you like, but do you really know how far you can reason without danger? Critique will say, in sum, that our freedom rides less on what we undertake with more or less courage than in the idea we ourselves have of our knowledge and its limits and that, consequently, instead of allowing another to say "obey," it is at this moment, when one will have made for oneself a sound idea of one's own knowledge, that one will be able to discover the principle of autonomy, and one will no longer hear the "obey"; or rather the "obey" will be founded on autonomy itself. [...] this true courage of knowing that was invoked by Aufklärung, this same courage of knowing [savoir] consists in recognizing the limits of knowledge [connaissance]; and it would be easy to show that for him autonomy is far from being opposed to obedience to sovereigns. But it no less remains that Kant affixed the understanding of knowledge to critique in his enterprise of desubjectification in relation to the game of power and truth, as a primordial task, as a prolegomena to any present and future Aufklärung.
>
> — Foucault 2019, 387

Critique, in this sense, is a way of recognising the current limits of human knowledge and the very potentiality of reaching beyond them. As for Kant (see 1784), it is a means of countering the "self-imposed immaturity" of humankind, an attempt to introduce autonomy into the process of subjectivation itself. A process of *desubjectification* is internal to any act of *enlightenment* under Power/Knowledge. However, if control is an increasingly pervasive way of applying discipline, if it is immanent to the processes that produce knowledge, meaning, and subjectivity, then the very space from which critique once operated faces the threat of becoming compromised. In disciplinary societies, the institutions that governed knowledge still had a discrete formation; although one cannot speak of an exteriority of power, there was nonetheless an *outside* to these institutions, a space between them, as Nathan Moore (2007, see above) names it. The configurations that could give rise to critique were more likely to form within this *outside*: although not free from the institutional grasp, a space where critique could still appeal to reason, truth, or moral law, partly beyond the institution. In control societies, however, the fluidity of dispositifs means that the partialities from which critique might emerge are scattered across the entire social field with less concentration. Power and knowledge coincide with the continuous modulation of information through the production of a specific rationality, but also through the *technologies* that can grind the information into the interiority of this specific rationality in an arguably more effective way. Critique risks being absorbed as another signal within the same feedback loop that (re)produces the regime of truth.

Having outlined how control generalises and intensifies disciplinary mechanisms, the question inevitably turns to the *Postscript*'s enigmatic closing section, *Program*. The notion of Program is often read as a double entendre, as the program of the mechanism of control and as a program for resistance (see e.g. MacKenzie and Porter 2021, 7-8). On the one hand, Deleuze presents a vague compilation of Félix Guattari's concept of cities from an unpublished screenplay, where access is regulated by codes and computational means (Deleuze 1992, 7),

where behind movement tracking and *machines saying* ***NO!***, by regulating access to facilities, control appears as the implementation of a gated society operating on dividual characteristics (e.g. biometric information on IDs). Deleuze also introduces a form of ressentiment as a consequence (or achievement) of control; both the corporate structure and the personalised ways of modulation put the subjects into a position to desire the continuous articulation of discipline. The never-ending training is now demanded by the subjects of this new societal setting (ibid., 7) in order to get ahead of others in the same class. This mechanism can be read similarly to the darkly satirical expression in Burroughs's (1979) "The Naked Lunch". His character, Dr. Benway, describes a mode of domination that operates not through overt force but by fostering guilt, diffuse anxiety, and the sense that subjection is deserved. Bureaucratic opacity completes the loop: the subject never encounters a clear agent of domination, only impersonal procedures.

> I deplore brutality [...] It's not efficient. On the other hand, prolonged mistreatment, short of physical violence, gives rise, when skillfully applied, to anxiety and a feeling of special guilt. A few rules or rather guiding principles are to be borne in mind. The subject must not realize that the mistreatment is a deliberate attack [...] on his personal identity. He must be made to feel that he deserves any treatment he receives because there is something (never specified) horribly wrong with him. The naked need of the control addicts must be decently covered by an arbitrary and intricate bureaucracy so that the subject cannot contact his enemy direct.
>
> — Dr. Benway (Burroughs 1979, 17)

Dr. Benway reads like a caricaturised product of the diagnosis Horkheimer and Adorno (2017) introduced in "Dialektik der Aufklärung"[11], namely the transformation of reason into an instrument of domination (or the instrumentality of reason itself). His clinical rationality serves no emancipatory end, only following a strict claim for *positive science*[12] leading to some grotesque and cruel practices throughout Burroughs' novels. However, as absurd as his introduction of the concept is, the ressentiment in control societies is precisely not just a more personalised correction of the subjectivity, but also the subject's wish to intensify the process in a competition with other members of the society.

On the other hand, in the sense of program as a program for resistance, something more unusual is taking place. Deleuze acknowledges the necessity of resistance and even suggests that collective formations such as unions may still retain strategic relevance (Deleuze 1992, 7), yet his analysis of the socio-technological machinery of control never develops into a prescriptive framework for how resistance could operate under these conditions. This ambiguity is surprising considering resistance is acknowledged as necessary, yet it is left without a concrete form. If control reorganises power into flexible, adaptive, and self-modulating forms, then resistance cannot rely on the same operational nature that was effective within the rigid enclosures of disciplinary institutions. The problem is not merely that control is pervasive; it transforms the terrain on which struggle unfolds. This gap has prompted later theorists to reconsider the question of resistance; Hardt (1998), for example, draws attention to how control operates within the broader dynamics of global imperialism, Galloway (2004) frames control as a protocol-driven environment that can only be challenged by intervening in its technical operations, MacKenzie and Porter (2021) emphasise the need for

[11] "Dialectic of Enlightenment"

[12] Derived from how Dr. Benway describes *pure science* himself:

> Balderdash, my boy... We're scientists. ...Pure scientists. Disinterested research and damned be him who cries 'Hold, too much !' Such people are no better than party poops.
>
> — Dr. Benway (Burroughs 1979, 66)

counter-sequential practices that interrupt the patterned chains through which power operates (see Section 2.4), while also stressing the role of critique in the Kantian sense of enabling an exit from an increasingly enclosing regime of truth (see below). A common theme between those works is that they agree on the immanence of the resistance to some degree; resistance is no longer imagined as a force standing outside power but as something that must take shape within the same infrastructures that organise contemporary forms of subjectivation, a theme that is either missing or not very well developed in the *Postscript*.

Deleuze's previous works, however, especially the ones in collaboration with Guattari (see "Anti-Oedipus" 1983 and "A Thousand Plateaus" 1987), are often read as pieces of theory that put resistance in the centre, since their concepts like "lines of flight"[13] and movements of "deterritorialisation"[14] are introduced as ontologically primary formations (Smith 2016, 278–280, see also Chapter 5 where Deleuze's work together with Guattari becomes especially relevant). By contrast, in Foucault's trajectory, the question of resistance emerges only late in his work, after his long elaboration of power's ubiquitous and constitutive nature. There are accounts in his later writings where Foucault insists that resistance must arise from within the very network of power relations that enclose and constitute subjects, and he even asserts its primacy and centrality to every power structure. Yet, as Daniel W. Smith (see ibid., 266) notes, this move is fraught with ambiguities: Foucault's resistance often appears reactive, secondary, and thus struggles to maintain an active, transformative quality as Deleuze also reflects on:

> It seems to me then that Michel [Foucault] encounters a problem which hasn't at all the same status for me. For if the systems of power are in some way constitutive, the only thing that can go against them are phenomena of "resistance", and the question bears on the status of these phenomena. [...] There is no problem for me in the status of phenomena of resistance: since the lines of flight are the primary determinations, since desire makes the social field function, it is rather the systems of power which, at the same time, find themselves produced by these assemblages [...] lines of flight, which is to say assemblages of desire, are not created by marginal elements [...] I thus have no need of a status of phenomena of resistance[.]
>
> — Deleuze 1997

Foucault's definition of resistance as an external phenomenon that occurs under certain conditions defines the phenomenon itself as prone to being easily reabsorbed into the structures it contests. In Gilles Deleuze & Felix Guattari (D&G)'s project, by contrast, power is inseparable from the investments of desire that compose social formations; resistance is therefore not exceptional or secondary but emanates from the same pillars that constitute the foundation of the power structure itself. The elements that consolidate a power formation are also those through which it can diverge, mutate, or collapse. Rather than locating an origin of resistance, therefore, D&G are concerned with understanding how immanent flows are formed into diverging, deterritorialising processes. Looking through their lens, a critical question emerges for the analysis of control: if *lines of flight* are indeed primary, if the formation of power is directly bound to the investments of desire that also constitute the forces of its deterritorialisation, and if Deleuzian and Deleuzoguattarian literature is so strongly characterised by its emphasis on resistance, then *why do lines of flight not emerge directly from the new*

[13] The Deleuzoguattarian term line of flight (ligne de fuite) refer to the formation that diverges from the established status-quo's grip, a path that enables parts of a system to break away, reconfigure, or deterritorialise existing structures of power, meaning, or order; "an infinitesimal possibility of escape" (Fournier 2014). A claim "that social formations are defined not by their internal contradictions, but by what escapes them" (Thornton and Holloway 2018, 14).

[14] Deterritorialisation is the process through which an established social or symbolic order (a territory) and its relations are altered, unbound, displaced. It "is the movement by which 'one' leaves the territory. It is the operation of the line of flight" (Deleuze and Guattari 1987, 672).

formations of power in control societies? Shouldn't the control as a novel formation also immediately cause new flows of deterritorialisation? And if it indeed does so, do these processes open any formulation of subjectivation against, through, or beyond control? Susanne Krasmann (2017) points out the issue and hardship in relation to the connection between the subject and power:

> Power brings the subject into being, but power does not exist independent of its enactment. It is immanent and only takes shape at a point of resistance. The subject is such a point of resistance that recasts, redirects and sometimes reverts power. Subjectivation, however, always involves wrestling with oneself; it is governing the self and self-government: the subject is bound to power as it is to him- or herself. How then to conceive of a political subject as a fold of power as well as a "line of flight"? How to imagine a challenge to the current regime of visibility?
>
> — Krasmann 2017, 18

Krasmann's insight displays the core difficulty in analysing contemporary forms of power. As the biopolitical mechanisms in the form described by Foucault and extended by Deleuze operate ever more closely upon the subject, the process of subjectivation becomes the very centre of power's formation. This proximity also renders resistance inseparable from the capacity to evade subjectification; the possibility of opening a line of flight becomes an irreducibly micropolitical task. Building such resistance, therefore, requires attention both to the machinery that produces subjectivity and to the practices of self-formation through which subjects participate in their own constitution. Where can we look for answers that Deleuze's *Postscript* leaves unresolved then? Where does subjectivation emerge today, when ubiquitous computing, surveillance, and granular knowledge of bodies shape the conditions of experience from the outset? Who or what can diverge from these dispositifs of subjectivation, and what forms can such divergence take? Should we once again look toward artistic or creative practices as potential sites of transformation? And how does this problem relate to generative systems such as LLMs, which increasingly participate in the production of meaning? Is their generativity merely reproductive, or does it contain possibilities that have yet to be explored?

So, what role remains for critical theory in a society of control? Can critique, conceived as "a source of understanding what we should resist, and how we should resist it" (I. Mackenzie 2018, 121), forge new paths beyond Deleuze's minimalist frame? Antoinette Rouvroy (2012, 13–14) reflects on it in the context of what she calls "algorithmic governmentality": "here, algorithmic regimes operate in a mode that bypasses confrontations with subjects, operating through infra-individual data and supra-individual profiles while masquerading as objective governance". As she notes, the experimental procedures of modern science are replaced by "real-time, pre-emptive production of algorithmic reality". This rationality assumes that data contains an objective truth that can be extracted with the right tools, and if it does not, the issue is framed as a lack of sufficient data. Although Rouvroy focuses primarily on decision-making architectures that are increasingly delegated to algorithmic processes, her diagnosis extends to contemporary genAI systems. Current tendencies in genAI development orient models toward ever-larger architectures and ever-expanding datasets in the pursuit of improved approximation to truth. Rouvroy's account, therefore, brings us back to the pressing question with a new look: in a pre-emptive and

micropolitical process of subjectification, where knowledge is extracted directly from data and prediction precedes interpretation, is critique still possible at all? Should we give up on it for good? One can read Iain Mackenzie's reflection as a response:

> The very conditions of control that shape our contemporary forms of governmentality are also those that enable immanent forms of resistance to control, because there is always the potential to switch around the direction of the 'IF...THEN...' functions in order to forge new connections. Even more importantly, though, there is always the potential to stall and disrupt these functions in the name of a 'what if'; the process of singularisation that accompanies the disruption of the algorithms themselves. [...] What is the art of resistance, today? It is the ability to disrupt the algorithmic flow of contemporary governmentality by connecting signs that don't function algorithmically; that is, subtracting the unique in the algorithm in order to form collective assemblages of 'what if' rather than 'IF...THEN...'.
>
> — I. Mackenzie 2018, 129-130

I. Mackenzie begins by insisting that critique must always be rethought in relation to contemporary forms of power; critique has a history, and it must adapt as each new social formation emerges. If critique seems impossible, it very well may be that the paths leading to the critique should be re-examined and reconfigured. For this purpose, his analysis turns to the strict definition of algorithms as "a self-contained step-by-step set of operations to be performed" (ibid., 125), in order to frame how contemporary computational logics condition the possibilities of political action. To develop a specific contrast among approaches towards critique, I. Mackenzie draws on Alain Badiou's criticism of Deleuze inspired politics. Badiou charges D&G with promoting a connectivist, rhizomatic[15] politics that mirrors the fluidity of neoliberal capitalism rather than resisting it. Yet, as I. Mackenzie argues, the formalism of Badiou's own approach is structured around an "IF... [event], THEN... [action]" schema that ends up resonating with the algorithmic logic it seeks to oppose (I. Mackenzie 2018, 126). In attempting to prescribe a politics grounded in fidelity to the event, Badiou risks reducing resistance to an algorithmic procedure: finite in scope, conditionally triggered, and ultimately compatible with the very computational regime he critiques. Referring to M. Lazzarato's (see 2014) notion of the sign[16] the fundamental operative units of algorithmic systems that pre-structure the social plane, I. Mackenzie (see ibid., 126–127) emphasises that resistance cannot follow the same procedural structure through which signs are formed in a regime. Algorithmic structures are strict and finite, whereas rhizomatics is a process-oriented way of binding signs in an infinite number of ways, which is what ensures that its emergence stays immanent and cannot be simply captured or subsumed by algorithmic governmentality (see ibid., 127-130). Rhizomatic composition thus enables an immanent critique that works through recomposing the very signs of control in ways that the system cannot fully anticipate or normalise. It is a way of keeping critique alive by opening new planes through connections that do not conform to the same pathways as control, which constitutes the precursor for resistance in I. Mackenzie's account.

Building on this foundation, he reminds us that "[c]ritique can no longer be conditioned by the reflexive subject able to determine the proper limits of the known, nor the transgressive subject able to go beyond the limits of the disciplines that es-

[15] A rhizome is D&G's concept for a non-hierarchical, non-linear form of organisation in which elements connect in multiple, shifting ways. Unlike tree-like or arboreal models based on roots, origins, and hierarchy, the rhizome figures thought, social formations, and practices as open, proliferating networks defined by connection, heterogeneity, multiplicity, rupture, and transformation (see Buchanan 2018, 622 for a short explanation and Deleuze and Guattari 1987 for D&G's introduction of the concept).

[16] Emanating from linguistics, signs as signifying elements are generalised by M. Lazzarato, who extends the concept beyond language to the semiotic operations of machines, objects, codes, and diagrams. As he writes, "signs (machines, objects, diagrams, etc.) constitute the focal points of proto-enunciation and proto-subjectivity" because "they suggest, enable, solicit, instigate, encourage, and prevent certain actions, thoughts, affects or promote others". More importantly, "through asignifying semiotics, machines communicate directly with other machines, entailing often unforeseeable and incalculable diagrammatic effects on the real" (ibid., 97). As I. Mackenzie elaborates:

> The sign can in principle take any 'computable' form: it may be a number, but it could just as legitimately be a visual symbol, a bodily gesture, a click on a keyboard, a smell; even a user's attentiveness or not to parts of a screen.
>
> — I. Mackenzie 2018, 125

tablish what is known" (ibid., 130). Rather, critique in post-disciplinary societies is about "opening up the possibilities" of divergence, alternative forms of subjectivation, connection, and collectivity; it must "embrace the processes implicit in algorithmic governmentality" (ibid., 130). It is mainly, therefore, I. Mackenzie turns to Guattari's exploration on art and aesthetics (see Guattari 1995a), the artist is first and foremost not an agent idly waiting to be activated by an event, nor her exploration follows "IFs" and "THENs". Artistic activity is for Guattari a rupture, an unframing that enables the creation of new individual and collective subjectivities. Guattari's artistic subject is an agent of process in exploration of new planes, of imagining, of generating new ideas (see I. Mackenzie 2018, 129), therefore, not just non-confirming with the algorithmic procedures but also more likely to bind the immanent partialities towards a push outside of it:

> [I]n rupture with signification and denotation - ordinary aesthetic categorisations lose a large part of their relevance. Reference to "free figuration," "abstraction," or "conceptualism" hardly matters! What is important is to know if a work leads effectively to a mutant production of enunciation. The focus of artistic activity always remains a surplus-value of subjectivity or, in other terms, the bringing to light of a negentropy at the heart of the banality of the environment - the consistency of subjectivity only being maintained by self-renewal through a minimal, individual or collective, resingularisation.
>
> — Guattari 1995a, 133

Guattari releases the artistic activity from the common associations; it is much less about appealing to some conventional aesthetic categorisations, and much more about a reconfiguration of *signs* in a way that it doesn't function algorithmically and forming the collective assemblages that built around what I. Mackenzie (2018, 130) calls "what if". Rupture in artistic pursuit is (not necessarily) a decomposition, rather an individual and collective reformation. And new subjectivations are not an external further pursuit; it is the other way around, the formation of new subjectivations generates artistic creation as a surplus. After all, from his perspective, the artist must "detach and deterritorialise a segment of the real in such a way as to make it play the role of a partial enunciator" (Guattari 1995a, 131).

The specific presentation of Guattari's artistic activity, with its emphasis on the reconfiguration of signs and the rupture of algorithmic processes, offers a concrete way of approaching resistance in control societies that most secondary literature lacks. Namely, the tendency to invoke lines of flight without specifying how such divergence might be technically or materially enacted within contemporary formations of dispositifs. In this context, I. Mackenzie delivers one of the most comprehensive accounts by both charting the connection between resistance and critique and framing their immanence to the contemporary infrastructures of power. Yet his reading rests on a characterisation of "algorithmic governmentality" as grounded in conditional and finite operations, a sequential logic that he, alongside Rouvroy, introduces without necessarily examining its current technical plausibility. Although their abstraction is mostly metaphorical, this becomes increasingly difficult to sustain once we consider the specifics of contemporary generative systems. It is a matter of debate whether the operation of these systems should be simply abstracted as "IF…THEN…" chains: their procedures are distributed, probabilistic, open-ended, and rooted in statistical in-

ference rather than symbolic rule execution, and such simplifications might create significant gaps in generalisation. Even if this abstraction holds as a loose metaphor, it risks telling only one side of the story; these models generate outputs in a continuous, flexible, and seemingly unbounded manner. It is also not entirely clear why the forms of resistance and critique cannot resemble the operational machinery of the control's dispositifs. Nevertheless, under such conditions, the immanent operation of critique and resistance, whether through a reconfiguration of signs or an interruption, could require the reconfiguration of the generative conditions themselves, a task that cannot be approached without analysing their underlying architectures. Furthermore, although the contrast between Badiou and the Deleuzoguattarian perspective establishes an important analytical frame, it remains unresolved whether divergence from the procedural pathways through which control operates can itself be taken as resistance, or whether following those pathways necessarily forecloses a possibility for divergence. This specific part of the argumentation calls not only for a close reading of the technical mechanisms at work, but also for an analysis of the institutional transformations that these mechanisms participate in, leading to even more questions than it reflects on.

2.4 *Towards a (Post-)Institutional Subjectification*

Framing critique and resistance on a new institutional transformation based on the technical aspects leads to a further discussion. If the institutional walls are breaking down and new technologies of power are taking over, the question becomes how the (post-)institutional framework of control societies can be read through the novel AI infrastructures. Can this framework account for the ways AI models govern, generate, and recombine information, and can it be linked, extended, or read in accordance with the architectures that increasingly mediate meaning-making? Algorithmic infrastructures imagined as barriers block flows, deny access, and enforce boundaries, but returning to Deleuze's provocation in the *Postscript*, the question "how can there be control if nothing is forbidden?" (Brusseau 2020, 2) is ever more relevant today. Contemporary growth of artificial models is much less about the control of access or prohibitions and much more about predictive analytics and pre-emptive adjustments.

While Deleuze's account remains foundational, its transliteral readings have led to an overstatement of the technological aspects and machinery, rendering control as a fully de-institutionalised form of power. Although from a temporal point of view, I. Mackenzie's earlier account appeared too early to engage with more sophisticated algorithmic systems and contemporary AI implementations, his later work with Robert Porter (see MacKenzie and Porter 2021) leans precisely into this issue. They emphasise that the transition from discipline to control must not be understood as the disappearance of institutions but as their transformation (ibid., 1–3). MacKenzie and Porter (ibid., 12–15), drawing on Erving Goffman's (see Goffman 1990) notion of "total institutions", characterise contemporary algorithmic formations as "totalizing institutions": disaggregated, permeable, and increasingly technical forms of authority that sequence dividuals across institutional domains. Here, totalisation does not mark a return to enclosed spaces but refers to the continual organisation of dividuated compon-

ents into malleable sequences through which institutional power now operates. While this formulation underscores the ongoing institutional character of control societies, it does not introduce conceptual commitments beyond what this chapter has already established. Building on this institutional analysis, however, "counter-sequencing" is introduced as a practical mode of resistance, specifically, as "the activity of reordering the power diagram of the totalizing institution in ways that destabilize its functioning" (MacKenzie and Porter 2021, 23). Rather than limiting critique to a process-oriented examination of algorithmic operations or to exposing the potentially infinite chains of signification constrained by procedural logics, counter-sequencing intervenes at the level of institutional ordering itself; an effort to reconfigure the arrangements that support and stabilise algorithmic governmentality. While counter-sequencing may involve nothing more than the disruption of a computational function, it can equally consist in *injecting* an alternative operational instruction that reorients the logic of a system, computational or otherwise. Such interruption does not promise an emancipatory or "positive" political outcome in advance; instead, it locates critical potential within the institutional logic of control, where even modest disruptions to its sequenced flows can open space for revaluation. As MacKenzie and Porter emphasise, the political value of counter-sequencing arises from its situated revaluation, not from any pre-given normative trajectory.

Yet, despite its conceptual promise, counter-sequencing remains only minimally developed; MacKenzie and Porter gesture toward its relevance for contemporary algorithmic and institutional formations, but they do not elaborate on how such reordering might be practised within infrastructures shaped by genAI. This is particularly striking because counter-sequencing appears far more directly applicable to the institutionalities and technical architectures examined throughout this chapter than the other discussed so far. The absence of a fuller account leaves open a question that becomes central for the present study: how can such interventions be understood once institutional sequencing is entwined with generative, anticipatory, and continuously recalibrating systems? It is at this point that the argument must widen again to its broader terrain, for this chapter has established several relevant points so far. The increasing convergence of the biopower to a pervasive micropolitical machinery in control societies put the subjectivation into the core of its formation and simultaneously rendered processes subjectivation the main ground for critique and resistance. If power and knowledge form a bi-equivocal relation, then any practice that opens new spaces of knowledge production immediately intervenes in processes of subjectivation and becomes entangled with the organisation of power itself. In this sense, critique cannot be reduced to a separate process, nor resistance to an exceptional intervention. They converge as a single immanent operation that acts where subjectivation is enacted. For this reason, taking a step further from I. Mackenzie's reading of critique as a precursor of resistance, I refer to this constellation as **Resistance/Critique**. Every attempt to open new planes of knowledge and meaning production not only reconfigures the conditions under which subjectivation takes place; it also generates new sites through which resistance can operate. Conversely, every interruption within existing sequences of power, however minor, creates further trajectories for thought, experimentation, and knowledge.

It becomes apparent that positioning Resistance/Critique in control societies in-

herently requires a methodological reorientation as well. It is no longer sufficient to speak of power in general; we must attend to the concrete dispositifs that composes the present. Understanding the machinery of contemporary control is essential to understanding the procedures of subjectivation. Although, as discussed above, it is in the literature not very well argued if and why Resistance/Critique should necessarily operate on a completely different procedural logic, this step is primary for any movement towards thinking about different forms of subjectivation. Moreover, as discussed earlier in this chapter, control presupposes not only technologies of modulation but also particular affective economies in which individuals strive to inhabit certain modes of subjecthood. Such dynamics build a form of ressentiment into the micropolitical condition. In this environment, the question of Resistance/Critique becomes inseparable from the question of cognitive entanglement with the infrastructures that *make* meaning. One must examine not only how power acts upon subjects but also how subjects participate in the reproduction of power by adopting its orientations. From this point on, every activity and every assemblage that reconfigures signs, as in the example of artistic pursuit, becomes particularly important for thinking human–machine interaction in the age of AI. To develop these claims further, the analysis now turns to the concrete terrain in which they unfold. These operations can be expressed as follows:

First to examine whether, and in what sense, the historical development and technical architecture of genAI models correspond to the (post-)institutional formations of control; in particular, the infrastructures through which meaning, behaviour, and subjectification are governed.

Second to assess how current debates and critical analyses of these models can be reframed when read through the technical account developed in this study, especially with regard to training processes, representational logics, and model behaviour.

Third to explore how resistance and critique might be (re-)configured under these conditions; whether they take shape as divergences, interruptions, or as emerging modes of subjectivation immanent to the very operations of genAI.

2.5 *Chapter 2 Summary*

This chapter traced the transition from disciplinary societies to control societies and analysed how this shift reorganises the production of subjectivity. Drawing on Gilles Deleuze (1992), it showed how enclosed institutions give way to diffuse and anticipatory mechanisms of modulation, increasingly mediated by computational infrastructures. These systems operate through continuous feedback, prediction, and dividuation, extending biopolitical mechanisms into a pervasive micropolitical machinery. The chapter revisited the genealogy of subjectivation to establish why, under these conditions, subjectivity becomes the central terrain on which power is exercised and contested, and why the analysis of subjectivation remains crucial for understanding contemporary dispositifs of control.

The chapter then examined the longstanding problem of critique and resistance within this environment. Through engagements with contemporary theorists in-

cluding Rouvroy (2012), Galloway (2004), and I. Mackenzie (2018), it highlighted the unresolved tension between the immanence of resistance and the increasing difficulty of articulating it under algorithmic forms of governance. Although the *Postscript* hints at the necessity of resistance, it leaves its concrete mechanisms undeveloped. Later work, especially MacKenzie and Porter (2021), introduced the concepts of totalising institutions and counter-sequencing, which together offer a preliminary framework for understanding how intervention might occur within computational infrastructures. These reflections culminated in proposing **Resistance/Critique** as a single immanent operation that emerges within the very procedures through which subjectivation is produced.

3

AI as the Infrastructure of Modulation

> I am less certain about treating machine learning as automation. Learning from data [...] often sidesteps and substitutes for existing ways of acting, and practices of control, and it thereby reconfigures human-machine differences. Yet the notion of automation does not capture well how this comes about. The programs that machine learners "write" are formulated as probabilistic models, as learned rules or association, and they generate predictive and classificatory statements ("this is a cat"). If this transformed calculability is automation, then we need to understand the specific contemporary reality of automation as it takes shape in machine learning. We cannot conduct critical enquiry into how calculation will automate future decisions without putting the notions of calculation and automation into question.
>
> Adrian Mackenzie 2017, 7-8

The previous chapter examined how the critique and resistance might be articulated under the institutional dynamics of control and its processes of subjectification. Management of information is central to those infrastructures and the Artificial Intelligence (AI) systems have an increasingly prominent role in the governance of information. Contemporary infrastructures analyse digital traces to generate personalised recommendations, assess relevance between users and content across search engines and social media platforms, and, with the emergence of Generative Artificial Intelligence (genAI) and Large Language Models (LLMs) (see Figure 3.1 for a guiding illustration to position different AI domains and products mentioned throughout the chapter), produce text, code, images, and other media. What were once predictive or classificatory instruments have become systems capable of synthesising and reorganising knowledge at scale. Their outputs now participate directly in communication, cultural production, and decision-making. Advances in Artificial Neural Network (NN) research and transformer architectures have been decisive in enabling these developments. GenAI models do not operate solely by identifying statistical regularities; they synthesise linguistic and visual material by drawing on extensive training corpora. Through such recombinations, they participate in the formation, circula-

tion, and reinterpretation of human knowledge. To understand how such architectures participate in the production of subjectivity, we must first trace the evolution of AI, from early symbolic reasoning to statistical modelling, Deep Learning (DL), and the emergence of self-attentive transformer architectures.

The present chapter focuses on the technical history and development of AI in order to understand how these models acquire their representational and generative capabilities in a chronological and conceptual methodology. It begins by outlining the trajectory of AI research, distinguishing between the symbolic paradigm (Symbolic Artificial Intelligence (symAI)) and the statistical approaches that underpin DL and genAI. This includes a discussion of NNs, Self-Supervised Learning (SSL), and transformers as the technical backbone of modern genAI models. Beyond description, this technical overview serves a strategic purpose: it shows how AI, even in its architectures, encodes specific logics of inference, representation, and control. These are not neutral design choices but material conditions that enable AI systems to act as infrastructures of knowledge production, decision-making, and governance. The chapter, therefore, provides both the technical foundations and the conceptual scaffolding for analysing the potentiality of genAI as a distributed, non-symbolic agent of control.

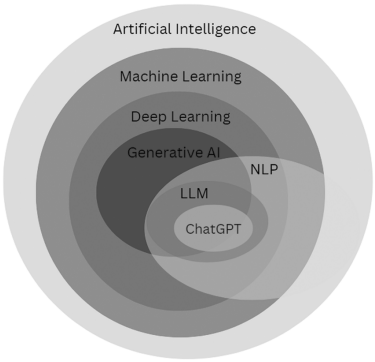


Figure 3.1: An illustration of overlap and interplay between AI domains leading to the LLMs such as ChatGPT (cf. Alomari 2024, 47)

3.1 *From Symbolic Rules to Statistical Inference: A Brief History of AI and Natural Language Processing (NLP)*

NLP is an area that lies at the intersection of linguistics, computer science, and AI, aiming to create computational systems that can interpret and handle human language data. It has been the ground for most of the breakthroughs in AI development, especially in recent years (see Bommasani et al. 2022, 22ff for a detailed analysis of the history of AI and language). Considering that, in some respects, the cognitive performance of an individual human is hardly superior to that of other primates (Manning 2022, 127), especially in short-term memory, it is hardly a surprise that the groundbreaking advancements in artificial pursuit of a mind happened on these shores. The transformative power of language has enabled *Homo sapiens* to link individual minds into collective networks of cognition. Language, rather than individual brainpower, constitutes the machinery through which human intelligence scales, distributes, and accumulates collectively (ibid., 127).

AI emerged in the mid-20th century, grounded in the formal logics of symbolic representation. The foundational paradigm, now referred to as symAI or Good old-fashioned AI (GOFAI), conceived intelligence as a matter of symbolic reasoning over explicitly encoded rules. The early paradigm treated intelligence as a computational process operating over discrete symbols according to explicitly programmed rules. AI systems under this logic were built to emulate deductive reasoning and problem-solving. The assumption was clear: if the world could be faithfully translated into a logical schema, machines could infer, deduce, and act rationally (see Eloff 2021, 183). Manning (2022) defines the **first era** between 1950 and 1969 as a development process under the immense lack of knowledge about the structure of human language or Machine Learning (ML) and AI. The 1956 Dartmouth Conference institutionalised the ambitions by defining AI as "the

science and engineering of making intelligent machines" (Montanari 2025, 195). Early research during this period was primarily focused on narrow, rule-based systems, particularly word-level translation lookups and simple mechanisms to handle inflectional forms and word order (Manning 2022, 128). In parallel, Alan Turing made substantial contributions by introducing the famous "Turing Test" (or "Imitation Game"), designed to evaluate a machine's ability to imitate human intelligence and rationality, along with the foundational concept of a universal machine (see Montanari 2025, 196). As Cognitive Robotics Prof. Murray Shanahan (see 2025) and Meta's Chief AI scientist Yann LeCun (see 2024) emphasise, the *Turing Test* is an inadequate benchmark for assessing modern AI models, but Turing's ideas nonetheless contributed to the conceptual foundation of the "prompt-based conversational machine" (Montanari 2025, 196). Aligned with Turing's perspective, the underlying notion in the early imaginary of a future AI was simple; if a machine could convincingly imitate a human in conversation, it would have been considered intelligent.

Relying on handcrafted rule sets meant implicit definitions of the features regarding the object of interest; for instance, to recognise patterns, the digit six in an image, one might encode the features "a closed loop at the bottom" and "a curve rising to the right". Such symbolic heuristics were sufficient so long as the data was clean and the context unambiguous. In the **second era** of AI development, spanning roughly 1970 to 1992, these approaches were extended to more complex domains, most notably natural language. By attempting to formalise aspects of linguistic structure and meaning, researchers pushed the boundaries of rule-based systems. While these models demonstrated greater sophistication in handling linguistic patterns, they still relied on explicitly encoded knowledge and remained limited by the inherent rigidity of symbolic architectures (Manning 2022, 129). Yet, these systems could not generalise beyond predefined rules. When confronted with noise or shifting contexts, their logic collapsed. The result was a period of stagnation and disillusionment now remembered as the "AI Winters" between 1970 - 1980 (Eloff 2021, 183).

But real-world ambiguity proved hostile to symbolic systems. As symAI attempted to scale into more complex domains like vision or language, it revealed its brittleness (ibid., 183–184). Philosophers of phenomenology were early critics of this paradigm, following Hubert Dreyfus' (2009)[1] earlier work where he argued that human intelligence was not symbolic, but embodied, situated, and fundamentally non-representational. Despite such critiques, symAI dominated the earlier decades of research in AI fields. This rationalist framework aligned with early cognitive science's attempts to model the mind as a rule-based machine of symbolic representation (see Montanari 2025, 194–197). Gilles Deleuze & Felix Guattari (D&G) were also one of the critiques, the hierarchically structured learning and the projection of a central pattern were clearly not working well:

[1] Originally published in 1972.

> This is evident in current problems in information science and computer science, which still cling to the oldest modes of thought in that they grant all power to a memory or central organ. Pierre Rosenstiehl and Jean Petitot, in a fine article denouncing "the imagery of command trees" (centered systems or hierarchical structures), note that "accepting the primacy of hierarchical structures amounts to giving arborescent structures privileged status.... The arborescent form admits of topo-

> logical explanation.... In a hierarchical system, an individual has only one active neighbor, his or her hierarchical superior.... The channels of transmission are preestablished: the arborescent system preexists the individual, who is integrated into it at an allotted place" (signifiance and subjectification).
>
> — Deleuze and Guattari 1987, 16

D&G's critique of early AI approaches centred on their rejection of hierarchical and centralised models, which constitute one of the main pillars of their project. Their affirmative alternative was grounded in a connectionist and non-hierarchical understanding of thought (ibid., 3ff.) as in opposition to arborescent or *tree* structures. While their critique targeted the symbolic, rule-based systems of their time, it is striking how closely their vision anticipated the architectural principles underpinning contemporary AI on a general level, particularly in its distributed, associative, and layered formations.[2] Nonetheless, the technological trajectory towards such architectures would take decades to materialise, revealing the prescient force of their philosophical intervention. The critique D&G (also) raised is arguably the foundation of the breakaway from the "IF...THEN..." logic in the future. Which is one of the circling analogical themes in the debate around algorithmic governmentality (see Section 2.3), where the analogy between programmable procedures and social regulation remains a central concern. Considering that the non-symbolic and connectivist structure of contemporary AI systems is one of the strongest motivations for this study to move beyond theories of algorithmic governmentality and reflections that still treat AI as if it operated through "IF…THEN…" architectures. It is also the reason for arguing that current debates can benefit from engaging D&G's broader work beyond the *Postscript*.

[2] Arguably, also regarding non-hierarchical functioning of the NNs; however, it is still a matter of discussion, if the genAI model architectures deploy a continuous subordination between different patterns and distributions. See the following sections for further discussion.

Following AI development aimed to overcome the shortcomings of symbolic sequences, and to find paths towards architectures without explicit definitions of instructions. The **third era** from roughly 1993 to 2012, was signified with the beginning of the abundance any novel AI innovation lacked the most, *the data*. As the internet boom suddenly introduced a massive digital corpus, researchers shifted toward statistical learning, leading to the rise of data-driven NLP. This shift replaced hand-coded rules with empirical models trained on annotated examples (Maas 2023); models could now generalise from data rather than deduce from explicitly defined axioms. Initially, the dominant approach centred on relatively simple statistical techniques applied to modest amounts of text, often in the low tens of millions of words. Researchers extracted linguistic facts from these corpora, identifying regularities such as common collocations or syntactic structures. Yet, early attempts to model language understanding through these means remained limited in their ability to capture deeper semantic or contextual knowledge (see Manning 2022, 129). For instance, early statistical models revealed that certain types of words tended to appear together, names of places often occurred alongside personal references, while more abstract terms exhibited distinctive distributional patterns. However, such surface-level regularities provided only limited insight into the deeper structures of language. As it became evident that simple frequency-based methods were insufficient for capturing the complexity of linguistic meaning, the focus shifted toward building annotated linguistic resources, such as syntactic treebanks, lexical databases, and labelled datasets for named entity recognition. These resources formed the foundation for more re-

liable, supervised learning approaches (see Manning 2022, 129). Onwards, the general purpose AI development continued with ups and downs in activity, with a couple of earlier successful neural network-based approaches like Mulloch-Pits. Among the early milestones was ELIZA, a rule-based program that mimicked a psychotherapist by matching keywords to scripted responses. Despite its simplicity, ELIZA gave the illusion of understanding and demonstrated the potential of machine conversation; though its developer emphasised it was merely parodic (Toloka 2023). Still, it signalled the beginning of natural language interaction with machines, laying the groundwork that statistical and later neural methods would build upon. Up until around 1997, much more advanced models like Deep Blue operating on more sophisticated architectures like the early attempts on Deep Artificial Neural Networks (DNNs) were developed (Montanari 2025, 197), but the main meta of the AI development was highly dependent on labelled data, and Supervised Learning (SL).

Although the real transformation originally began in the early 2000s, the first significant fruits of the new direction dropped around 2013, which marks the **4. and current era** in AI development (Manning 2022, 129). Pushing through the ability to process more and more data allowed a new paradigm to emerge, rooted in NNs inspired by the architecture of the brain, *connectionism* became the new meta of further advancements. These systems, now more broadly applied and clearly defined as DNNs, learned not by logic but by adjusting distributed weightings across layered networks, which became the foundation for contemporary ML and DL systems. Exponential advances in computation enabled these networks to scale (Eloff 2021, 184) and finally also pushed towards an Unsupervised Learning (UL) methodologies, whereas the models were geared towards recognising patterns in the data without being explicitly told which features of the data were pointing to what. For instance, while early augmentational models were trying to distinct between cat and dog photos by looking at photos labeled by humans and other processes as either as *dogs* or *cats*, UL models are looking at a data collection of unlabeled photos and try to find patterns in them which makes both parties distinct through specific characteristics, in other words, towards finding out about the substance of *dogness* and *catness*. On the NLP fronts, linguistic units such as words or sentences came to be represented as vectors in high-dimensional vector spaces. Semantic and syntactic relationships were modelled not through rule-based analysis and pre-defined categories, but through the spatial proximity of these vectors (Manning 2022, 129). DL allowed to parse distant context, as well as processing the words meaningwise close to each other, thanks to this generalised vector space approach optimised with more and more textual data (see ibid., 129). This approach turned out to be far more effective than earlier attempts at formalising linguistic meaning. Instead of hand-coding grammatical rules or manually annotating small corpora, models could now process large textual datasets and infer structure statistically. DL enabled systems to capture long-range dependencies in context and identify meaning-level relationships through learned representations optimised across massive datasets. Crucially, this reduced the need for manual labelling, as UL techniques became dominant.

One of the most significant turning points was around 2018 with the successful implementation of the SSL approach. SSL constitutes a special case of the UL,

which not only makes the models identify underlying structures in the data but also enables them to create their own training exercises through the prediction challenges they are subjected to (ibid., 129). This includes masking specific words in the text to try to predict the correct or most fitting tokens, or try to guess the next word in an abruptly cut text, where SSL models learn by predicting missing elements from within the input itself. This method allowed models to learn linguistic regularities from massive unlabeled corpora, and it gave rise to pre-trained genAI models (Maas 2023). The novelty that specifically enabled this leap was the *transformer architecture*. Its core mechanism, self-attention, computes weighted dependencies between all tokens in a sequence, allowing the model to capture long-range relations independent of word order. This innovation enabled massive parallelisation and scalability (ibid.). Availability of vast data and the unique novelty of transformer architecture that was powered by a huge amount of reinforcement capability through repetition has been crucial in operating on SSL methodology to parse and accumulate huge amounts of unlabeled human language data.

3.2 *Mayan Codices and Telephatic Broadcasts: Algorithmic Governance of Information before GenAI*

From William S. Burroughs's (1979, 81) "The Naked Lunch", a relevant quote can be found below.

The earlier AI implementations on the web are mainly classified as recommender systems, which associate relations between different content, and filter accordingly. Their widely still relevant application has started with the participatory internet culture, where users also became contributors, for example, on social media platforms. Krassmann notes that this transition rendered humans rapid data generators for the training sets:

> Thus far, we have determined that whereas the individual and disciplinary power seem to be cast in the same mold – the former being the product of the latter – the digital subject of the control society 2.0 appears to be an active subject able to make decisions – which in turn feeds the algorithms.
>
> — Krasmann 2017, 19

This insight offers a precise entry point into the history of AI. Long before the emergence of genAI, NN based AI systems were integrated into infrastructures designed to sort, rank, and anticipate behaviour. Search engines, recommender systems, and ranking algorithms constructed profiles, inferred preferences, and organised interactions through relevance estimation (see Demir 2019, 26–30). These systems already relied on a feedback-driven logic: user behaviour shaped algorithmic output, and algorithmic output shaped subsequent behaviour. When read through Deleuze's diagram, such early NN applications exhibit the operational dynamics of control; continuous capture, iterative adjustment, and subtle steering of conduct. Their core mode of operation can be summarised in the following loop:

1. massive data collection from user interactions,
2. indexing and probabilistic categorisation of behaviours,
3. ranking and recommending content based on **relevance association**,

4. generating personalised information flows, recommendations, associations,
5. feeding back the gathered information into the user's profile to update the personalised process (see Figure 3.2 for an illustration of the process).

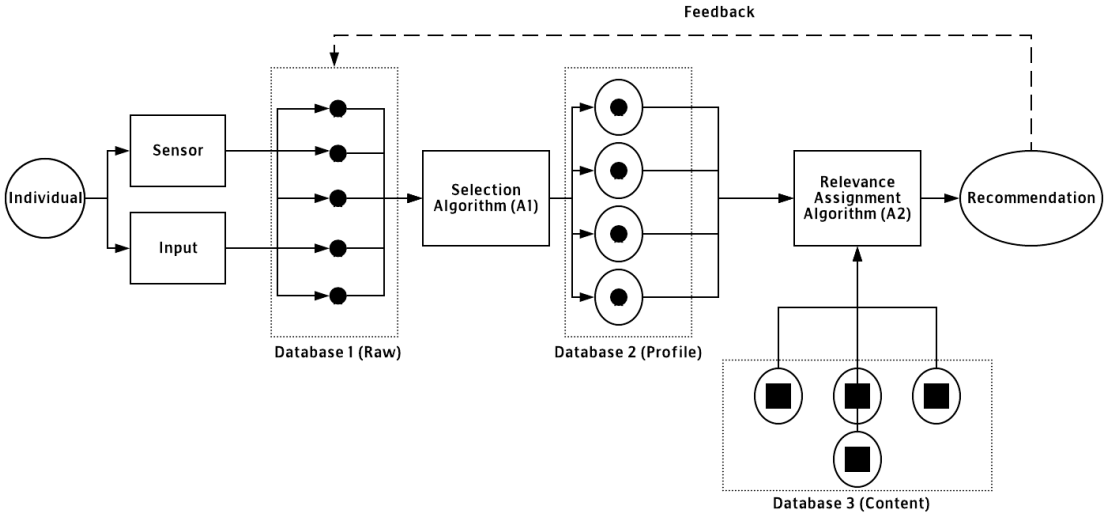


Figure 3.2: Algorithmic Selection and Relevance Assignment Process (cf. Just and Latzer 2017, 241)

The anchoring process produced by the feedback loops between users and recommender systems (see the characterisation of anchors and endless while-loops in Demir 2019, 34–35) established a correction mechanism based on the dividual traces of users, that is, the fragments of data assembled by algorithmic systems to construct *profiles*. Each interaction became an input to a probabilistic model that then shaped the horizon of the next interaction. Platforms such as Facebook or YouTube did not need to coerce users; they governed behaviour through *environmental modulation*, subtly reinforcing predictable patterns of attention and engagement (ibid., 29–32). Yuk Hui names it as a process of "disindividuation":

> Under the guise of being free and friendly to use, we can see in this example that the modulation of social relations can actually lead to what we have called 'disindividuation' [...] the attention of each social atom (or 'person') is sliced into ever smaller pieces and dispersed across networks via status updates, interactions, and advertisements. [...] The 'collective' on Facebook becomes a distraction, a cause of the dissolution of structures within individuals, but not a site of new modes of empowerment.
>
> — Hui 2015, 90

The recommendation systems and the algorithmic governance of information reflect the process of dividualisation that characterises control societies (see also Cheney-Lippold 2011); the coherence of personal or collective agency is fragmented into algorithmically analysed micro-traces of digital history. While this dynamic dissolves the unitary subject, the constant personalisation of digital reality enacts what Deleuze calls modulation: an ongoing, fine-grained adjustment of the individual's field of experience. From an infrastructural perspective, these algorithms already governed information flows and the subjectification process,

creating a precondition for the transition to generative systems. Whereas these early models merely filtered, ranked, and nudged, contemporary genAI systems will move beyond governance of information toward its generation.

However, returning to the question of the nature of *control*, it is worth asking whether the institutional mechanisms of control were ever meant to be generative in the first place? Were the computational methods associated with control societies ever intended to communicate with individuals rather than simply act upon their traces? And is adaptation, flexibility, and the articulation of statistical inference sufficient to classify genAI systems as dispositifs of control? If critique and resistance are to be reconsidered under these conditions, the novelty introduced by generativity becomes a central concern. What does it mean for systems of control to produce, generate, and respond, rather than only to filter and anticipate? As a satirical analogy for the limits of what constitutes control, Burroughs offers a definition of biocontrol in "The Naked Lunch":

> The biocontrol apparatus is prototype of one-way telepathic control. The subject could be rendered susceptible to the transmitter by drugs or other processing without installing any apparatus. Ultimately the Senders will use telepathic transmitting exclusively... Ever dig the Mayan codices? I figure it like this: the priests – about one per cent of population – made with one-way telepathic broadcasts instructing the workers what to feel and when... A telepathic sender has to send all the time. He can never receive, because if he receives that means someone else has feelings of his own could louse up his continuity. The sender has to send all the time [...]
>
> — Burroughs 1979, 81

One can read Burroughs's description as a useful contrast for distinguishing generative systems from earlier applications of AI. At first glance, the early NN-driven platforms already align with the institutional description of control societies. They dissolved enclosures, operated through environmental cues, and extracted dividual traces from users; this constituted the paradigm of *algorithmic governance of information*. The emergence of genAI creates a different constellation. These models do not simply modulate existing flows of information but *generate* content, narratives, and knowledge formations that participate directly in the shaping of subjectivity; the machinery of governance, therefore, becomes a machinery of production. I argue that while these systems maintain a strong resemblance to the dispositifs of control, their generative capacity introduces a degree of novelty for thinking about critique and resistance. Whether this development extends the logic of control or marks a qualitatively distinct mode of operation leads to the following tasks:

1. to open the black box of genAI and its transformer-based architecture;
2. to examine how these models mediate human agency and the production of meaning.

3.3 *Deep Learning (DL) and Generative Artificial Intelligence (genAI)*

At their core, contemporary generative systems are NNs. A NN is a computational architecture inspired (loosely) by biological neurons, each neuron receives

inputs, applies weights and biases, passes the result through an activation function, and transmits the signal forward (for one of the fundamental papers, see Rosenblatt 1958). What distinguishes NNs from earlier symbolic systems is not rule-following but function approximation. By adjusting millions or even billions of parameters during training, these architectures learn statistical mappings between inputs and outputs that cannot be written down as explicit rules (see LeCun et al. 2015; Rumelhart et al. 1986). DL extends this principle by stacking many such layers. Depth allows the network to build hierarchical representations: lower layers detect relatively simple features, while higher layers capture progressively abstract patterns (see Figure 3.3 for a simple illustration). Instead of storing meaning in explicit symbols, meaning emerges from distributed patterns of activation spread across the network. This is what enables the modelling of highly non-linear relationships in data, crucial for handling the complexity of natural language, vision, and multimodal inputs (see e.g. Goodfellow et al. 2016; Schmidhuber 2015).

GenAI models, particularly LLMs, are thus best understood as specialised deep NNs. Instead of operating on predefined linguistic rules, they function by encoding massive distributions of textual patterns into weight configurations. Their generativity stems from this architecture; by sampling from learned distributions, they produce novel outputs aligned with the statistical structure of language. In this sense, the architecture itself is the key to their meaning-making capacities, their seemingly impressive way of binding distant concepts.

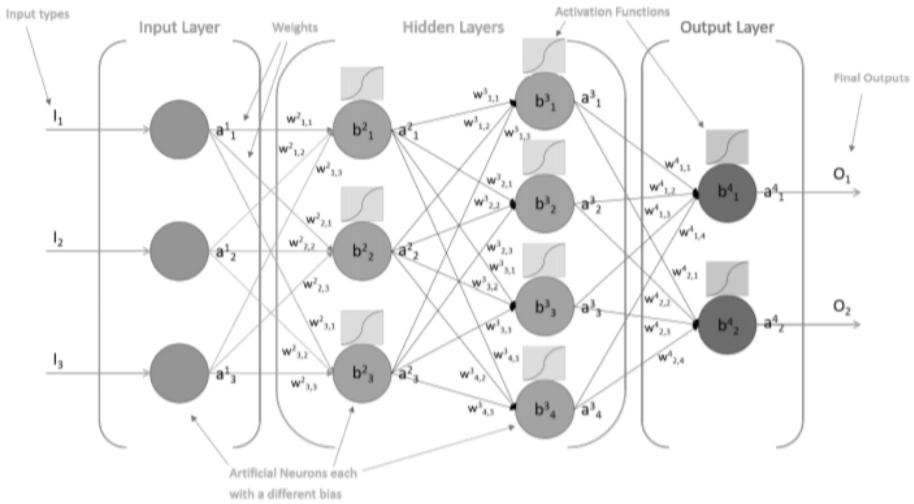


Figure 3.3: A Simplified Illustration of a NN (cf. Subramaniam and Kaur 2019)

But how does the generativity function? How, in fact, is meaning produced? Does the machinery itself offer clues about the nature of the content that genAI models generate? Beyond the corporations that develop the most sophisticated models, beyond the deliberately added specific configurations, there lies a common architecture animating these meaning-making systems. What remains, if not to look directly into the machine? Whether because of its complexity or a lack of potential for insights, this path is an especially unexplored one in the critical theories of the contemporary sociotechnological advancements. Following the discussion in the last chapter, exactly because of this specific lack, I am delving into an analysis of the specific features of the genAI models by specific-

ally focusing on how **a transformed based LLM** produces outputs.

3.3.1 *Vector Spaces and Collapsing Dimensions*

After the breakthroughs in NN architectures over the past decade (as outlined in Section 3.1), many influential designs for sequence modelling, particularly in machine translation and NLP, were based on Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). Despite their advances over earlier implementations, these models faced a fundamental limitation often described as the problem of *locality*: their difficulty in capturing long-range dependencies across sequences (see e.g. Bengio et al. 1994). In ML, data must be reshaped into a form the model can process. For NLP, this requires vectorising language into a high-dimensional space where tokens are assigned coordinates and scales.[3] Once embedded, relations between elements can be computed algebraically, allowing the model to operate within what A. Mackenzie[4] (2017, 51) calls an "expanding epistemic space", where results emerge from geometric proximity and transformation.

[3] Since a word is the most common form of a token in NLP, vectorisation means representing it as a vector $(x_1, \ldots, x_n)$, where each component x_i corresponds to a dimension in the embedding space. The number of dimensions n is fixed by the model's architecture and determines how tokens can be compared and transformed. For instance, common embeddings use $n = 300$ dimensions in *word2vec* or $n = 768$ in *BERT* (see Mikolov et al. 2013).

[4] Since two different authors with the last name MacKenzie are cited in this paper, note the distinction between Iain Mackenzie, cited as I. Mackenzie, and Adrian Mackenzie, cited as A. Mackenzie.

Once vectorisation is performed, the next question is how to most effectively represent the resulting vector space. In its raw form, data often contains a very large number of features, which translate into dimensions that are too burdensome for models to handle directly. This motivates the development of *dimensionality reduction* techniques. Far predating the rise of genAI, dimensionality reduction is a foundational method in ML that projects high-dimensional data, such as raw image pixels or token embeddings, into a compressed latent space that is more tractable for statistical operations prior to training. These latent representations are not merely a technical convenience; they constitute the terrain upon which inference, generalisation, and generation take place. In this process, each data object, whether a sentence, an image, or a behavioural trace, is mapped onto a point or trajectory within a lower-dimensional space. The resulting representations emphasise the most *distinctive* features relevant to the dataset as a whole. In the contemporary sophisticated DL models, analogous forms of dimensionality reduction occur within the intermediate layers of the network, since training requires the data to be represented at different levels of abstraction; these transformations do not necessarily reduce dimensionality and can at times expand it, but they nonetheless perform a comparable compressive or restructuring function, and to illustrate this more intuitively we can turn to earlier, more explicit implementations of dimensionality reduction in classical ML. Indeed, dimensionality reduction methods such as Principal Component Analysis (PCA) are often used to "flatten the vector space down into lower-dimensional subspaces" (ibid., 73). This approach reduces complexity, highlights dominant patterns, and improves the efficiency of subsequent learning tasks (see e.g. Jolliffe 2002, 1–9) with a trade-off of losing some information from the initial raw data. However, dimensionality reduction necessarily involves choices about which aspects of the data are preserved and which are discarded, and this selective compression underlies concerns about the representations that genAI models construct, since they are grounded in a reduced and fundamentally *latent* reality (see Chapter 4.1).

Dimensionality reduction might be hard to visualise in the case of text data, but image recognition models often deliver better insight into the operation. See the

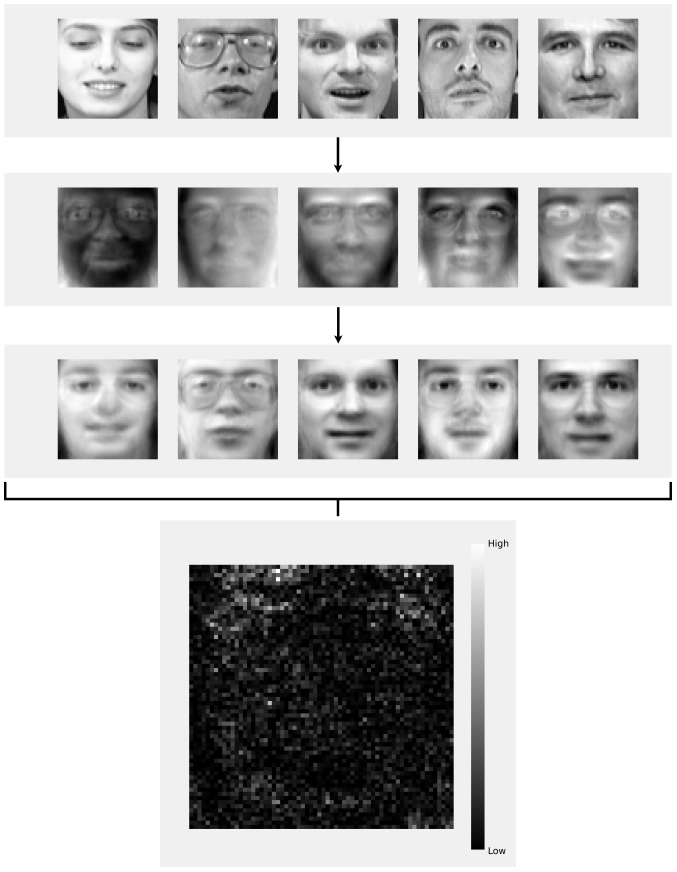


Figure 3.4: Dimensionality Reduction via Principal Component Analysis, Image Reconstruction out of 20 Principal Components, and Feature Importance Visualisation using Olivetti Faces Dataset (dataset: ATT Laboratories Cambridge 2005, implementation: author's self work, see Annex A.)

example in Figure 3.3.1: it shows snapshots from the training of a simple image recognition model on the Olivetti Faces dataset (see ATT Laboratories Cambridge 2005; a collection of standardised, grayscale portraits of 40 individuals). On the first row, there are portraits of different subjects in the dataset. On the second row, we see five random principal components obtained from the dimensionality reduction operation via PCA[5], which can be thought of as the building blocks the model uses to (re)construct faces in a more compact representation. In the third row of the figure, we see the same five faces from the first row, but reconstructed using 20 principal components extracted in the PCA process.[6] The reconstruction is shaped by the stronger features across the dataset: the new faces blend the traits that make faces distinctive, while highly individual features survive only insofar as they represent strong divergences in the dataset as a whole. See, for example, how the reconstructed images also contain features from other faces; one of the most distinctive examples of this is that all of the reconstructed images now feature some resemblance to glasses. The reconstruction is a reimagination of faces by using the most distinctive aspects of all of the faces.

Finally, the last visualisation displays the locations of the most important features for the model on each pixel, with lighter pixels indicating higher importance. As one can observe by looking at the lightest pixels, these are the most emphasised parts of the reconstructed images while still being the parts where some *ghost* features (like the outlines of the glasses example) blend the most. The model is therefore more likely to preserve and emphasise those features that play a distinctive role across samples. In other words, the registration of the data is trans-

[5] On the mathematical level, these correspond to the *eigenvectors* of the sample covariance matrix $X^T X$. Each eigenvector points in the direction of maximal variance, and the associated eigenvalue measures the strength of that variance.

[6] This reconstruction through 20 principal components corresponds to an explained variance of roughly 70% (see Annex A), meaning that the majority of the dataset's information content is retained even after compression.

formed through a reconstruction guided by these selected features. Although this demonstration is based on images, the same logic underlies dimensionality reduction in genAI contexts. When applied to text, the "faces" become words and contexts, and the principal components become latent dimensions of meaning. Just as the Olivetti reconstructions compress facial features into a tractable subspace, LLMs compress linguistic variation into latent vectors, privileging what is most statistically distinctive while likely discarding subtle or marginal patterns. This illustrates how latent representations, whether of faces or words, are always a reduced lens on reality; efficient and powerful, but partial.

The logic of dimensionality reduction illustrates how high-dimensional data can be compressed into tractable latent spaces that retain the most distinctive features of a dataset. In the domain of sequence modelling, this logic was taken up in early neural architectures such as RNNs and CNNs, which typically followed an encoder–decoder design. These architectures operationalised the principle of latent representation: the encoder compressed an input sequence into a continuous vector space, and the decoder expanded this representation into an output sequence.[7] RNNs processed tokens sequentially, passing information through hidden states that decayed over distance, which made capturing long-range dependencies difficult. While they could build semantic connections reasonably well, they failed to construct robust models of language regardless of the training scale. CNNs, while more parallelisable, were constrained by kernel sizes and fixed receptive fields. Both designs struggled with tasks requiring global relational awareness of a sequence. In order to both build the long-distance relationships layered over huge datasets, and internalise the play between dimensionality reduction and reconstruction on a multi-processed surface, something much more powerful was needed.

[7] In its simplest form, the encoder takes an input sequence of symbols $(x_1, \ldots, x_n)$ and transforms them into a sequence of continuous vector representations $\mathbf{z} = (z_1, \ldots, z_n)$. These vectors encode the relevant information from the input. The decoder then generates an output sequence $(y_1, \ldots, y_m)$ one step at a time. It is *auto-regressive*, meaning it uses previously generated outputs (e.g. $y_1, y_2, \ldots$) as input when generating the next token. This setup allows the model to generate coherent and context-sensitive output, building each element of the sequence in a structured, history-aware manner (Vaswani et al. 2017, 2).

3.3.2 *Transformative Attention and Signs without Signification*

The Transformer architecture marked a decisive break from the sequential bottlenecks. Dispensing with recurrence and localised convolution, it introduced *self-attention* as a mechanism for computing contextual representations. In simple terms, self-attention allows the model to decide which parts of the input are most relevant to each other when producing an output. In a single operation, every token in the input sequence attends to all others, producing weighted combinations of contextually relevant elements (ibid., 4). In their groundbreaking paper, "Attention Is All You Need", Vaswani et al. (ibid.) proposed a new architecture that preserved the encoder–decoder structure but eliminated reliance on recurrence and convolution. Instead, the Transformer model relied entirely on attention mechanisms, not as a supplementary feature but as the foundation of both the encoder and the decoder (ibid., 1–2; see Figure 3.5 for an illustration). This architectural shift allowed for highly parallelised computation, better modelling of long-range dependencies, and significant improvements in scalability. The Transformer has since become the cornerstone of contemporary genAI, enabling many of the recent breakthroughs in large-scale language modelling and generative systems. The architecture is built from stacked encoder and decoder layers, each composed of multi-head self-attention and pointwise feed-forward networks. These attention heads act as differentiated channels through which the model

adjusts its internal representations, integrating multiple semantic and syntactic perspectives concurrently. Instead of treating tokens as isolated or sequential entities, attention turns the entire sequence into a site of mutual interaction, where each token is redefined in relation to all others. By eliminating recurrence and convolution in favour of attention, the Transformer achieved two decisive outcomes: first, it enabled much more comprehensive and effective training on vast datasets; second, it allowed the model to capture long-distance connections and complex contextual relations with unprecedented efficiency, thereby overcoming the failure of previous NN architectures to produce a representation capable of capturing the essence of the vast datasets on which they were trained. These properties form the *technical substrate* upon which modern genAI and LLMs are built, leading to text-to-text models like ChatGPT, as a result of developments in NLP, and models like Midjourney in image understanding and computer vision (see Ploennigs and Berger 2023, 2).

Conceptually, the Transformer establishes a *global field of relation*, where each token is encoded not in isolation or rigid sequence but through its distributed relevance to all others. This process builds on the algebraic nature of the tokenised dataset embedded in the high-dimensional *vector space* (the feature space introduced above), where semantic and syntactic relationships are captured as measurable distances. The architecture thereby creates a form of synchronic awareness: the presence of every other word is embedded within the representation of each word. The high-dimensional *feature space* encodes tokens as points separated by specific distances (as a metric space), turning both the position of a token and its relation with other tokens into numerical values, making it possible to perform relational operations such as $king - man + woman \approx queen$ (AIG 2025b).[8] Similarly, a polysemous word such as *bank* can be shifted towards different meanings: if the position is shifted in the direction of finance, its neighbourhood becomes populated with tokens such as *securities, banking, investment, credit*, whereas if shifted towards the position of *river*, its neighbourhood changes to *flows, shore, stream, along*, and so forth (see ANNEX B for a demonstration). The same mechanism, however, can also embed and even amplify specific biases in the data. For instance, embeddings may position professions such as *director, officer, policymaker, programmer* closer to male-related tokens, while *hairstylist, receptionist, nurse, veterinarian* are drawn towards female-related tokens. Likewise, scientific terms tend to cluster more closely with male-related tokens, whereas terms linked to the arts are positioned nearer to female-related ones (see ANNEX B for the code and examples).

[8] See a demonstration of this well-known operation in the *GloVe* word vector space Pennington et al. 2014, presented in ANNEX B. The operation yields a cosine similarity of 0.861 (very high) for the terms on both sides of the equation.

This reconfiguration of relationality is the basis of the efficiency, scalability, and generative fluency that define modern LLMs[9]. Modern architectures, however, go well beyond this initial plane of departure. One of the most important aspects that gave transformer-based systems their edge over anything else in terms of relationality was how *attention* was utilised. The *attention mechanism* is mainly responsible for improving the interaction between input and output, allowing the model to dynamically focus on the most relevant parts of the input sequence while generating each token. Attention computes a set of weights over the input representations, effectively answering the question: "Which parts of the input matter most for predicting the next output?"[10] Each token's final representation is thus a weighted blend of all other tokens, adjusted by their contextual

[9] Although we are focusing specifically on LLMs here, the transformer architecture is also embedded in other genAI models such as *text-to-image* generators.

[10] Technically, self-attention calculates relationships between tokens by projecting them into *query*, *key*, and *value* vectors. These are used to compute attention weights through dot-product similarity and softmax normalisation.

relevance. Through multiple stacked layers and attention heads (multi-head attention), the transformer operates on different planes of relevance. For example, while one head may attend to the single most relevant token in the input, others simultaneously track secondary relations or longer-range dependencies. In this way, multiple assessments of relevance are carried out at once, both within the input sequence itself and between the input and the model's internalised representation of the whole training data (see e.g. Merritt 2022). As Vaswani puts it (ibid.), "meaning is a result of relationships between things, and self-attention is a general way of learning relationships." Probabilistic modelling, together with these long-distance relational adjustments, then governs how the network moves through representational space to predict the next output (see Montanari 2025, 198). Moreover, the multi-head parallelisation of attention attached to both encoder and decoder processes (see Figure 3.5 for the official illustration) "allows the model to jointly attend to information from different representation subspaces at different positions" (Vaswani et al. 2017, 4). The transformer model, so to speak, has "*radicalised* the use of attention in sequence-to-sequence language modelling, dispensing entirely with recurrence and convolution in favour of an ensemble of attention mechanisms" (Amoore et al. 2024, 6).

Maas (2023) associates this novel operational structure of the Transformers with Derrida's concept of *trace* (see e.g. Derrida 1998, 26). Derrida's concept is an advancement of Ferdinand Saussure's linguistic theory of *signifiers* and *signifieds* (see e.g. 2007) through his own concept of *différance*, in which the emphasis shifts to the context-dependency of words and their differentiation from each other. For instance, the colour *red* is defined through its differentiation from *green* and *blue* without having any actual substance of its own (Maas 2023, 9). "The sign has no component that belongs to itself only; it is merely a collection of the traces of every other sign running through it" (Cilliers 2002, 44). All signs are in continuous relationship with other signs, where the position of a word within the current network of connected signs, and their differences[11] from that specific sign, establish its substance. Yet the substance or *meaning* of the sign has temporal dependency, because the specific arrangement of words, as well as the differentiation between them, is in constant flux, "in a dynamic process of combination and referencing" (ibid., 44), dependent on the current context[12]. Similarly, in the operation of LLMs, this spectral interdependence, where tokens are mutually inscribed into one another, suggests a structure in which meaning is always already haunted by the rest of the utterance (see Maas 2023, 12) in the sense of Derrida's *trace*. The *meaning* of words in LLMs is defined by overlapping distributions: the distribution within the sentence, the distribution the model renders across the whole dataset, and other dynamic mechanisms regulated by the Transformer core all working on the signification layers in the formation of traces going through the specific word (token) attention mechanism is focusing on.

Montanari (2025) draws a direct connection between the cognitive functions mimicked by Transformer architectures and the cultural implications of genAI models. The ability to construct relationships between concepts that are distant from one another is precisely what enables LLMs to understand and articulate metaphors[13].

[T]ransformer models, which exemplify the interplay between metaphor and func-

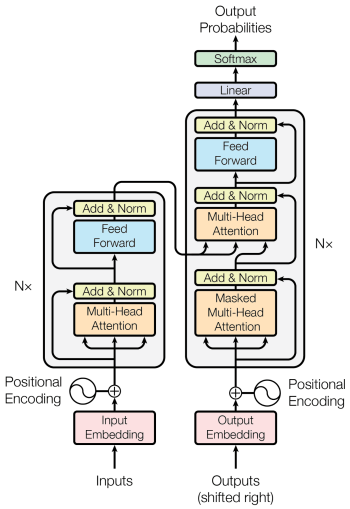


Figure 3.5: The original Transformer Architecture with built-in Multi-Head Attention Mechanism in Encoder and Decoder Processes (cf. Vaswani et al. 2017, 3)

[11] In terms of word embeddings in LLMs, we can interpret these differences as distances, since distances in the network represent the model's way of encoding differences between concepts.

[12] I will be referencing this temporal formation as an *instance* from now on, since the context-dependency of the network narrows meaning into one instance of the connectional structure.

[13] For a short and engaging discussion of this capability, see Heerden and Bas (2024), who trained a simpler form of LLM that nonetheless succeeded in grasping and generating poetic metaphors in a low-resource language such as Afrikaans, using only a fraction of the text data available for English.

> tion. Transformers [...] simulate certain structures and functions of the human brain, excelling at processing sequential data such as words in a sentence or notes in a melody. The transformative innovation within Transformers is the "attention mechanism," which enables the model to focus selectively on the most relevant parts of the input sequence. This mechanism is pivotal for discerning complex relationships and dependencies within data. [...] multi-head attention mechanism, a key feature that captures diverse aspects of an input sequence simultaneously. This dual role of technical objects – functionally specific and mythically resonant – reveals their broader cultural impact. Technical metaphors, often catachrestic and hybridised, solidify not only the utility but also the mystique and credibility of AI systems.
>
> — Montanari 2025, 206

Montanari's analogy between metaphor and function illustrates how the distinctive capacities of transformer architectures become visible at the level of their outputs. By design, transformers are highly efficient translation machines. One of the most prominent challenges in NLP that the transformer architecture immediately rendered trivial was language translation. Yet this capacity extends beyond linguistic translation: the same mechanism of associating distributions across data allows for effective cross-modal mappings, such as text-to-speech or text-to-image generation. From the perspective of meaning-making, the production of sense in transformer-based models can be understood as a continual translation, moving between stratified elements and overarching concepts, where meaning emerges fluidly from the situated application of traces within each exchange. AIG (2025a) points to the D&G's notion of "double articulation" as a way of theorising this machinery. Double articulation concept in D&G's theory describes how structures are formed on two surfaces of production: a molecular articulation, where raw flows of matter, energy, or desire are segmented, and a molar articulation, where these segments are organised into larger social, linguistic, or institutional forms. For example, in language, sounds (molecular) are articulated into words and grammar (molar). This shows how every stratum, from biology to society, emerges by combining micro-processes with macro-organisation (see e.g. Chapter 3 in Deleuze and Guattari 1987).

The Transformer operates simultaneously on two strata: a molecular level of local attention, where data is tokenised, and neural activations are formalised into specific connection patterns that correspond to certain concepts, clusters in the feature space, or relations of distances and neighbourhoods; and a molar level, where these are aggregated into larger representations and models capable of generation. These molar structures regulate flows of input and output, steering and shaping responses (see AIG 2025a). In the language of D&G, every input sentence first undergoes a process of deterritorialisation[14], where its components are broken down, only to be reterritorialised according to the molar aggregates the model has constructed in order to generate a response. AIG provides an exemplary process:

> Consider the example of processing the sentence "She is a scientist. She conducted an experiment":
>
> 1. Each token [in input] ("she," "is," "scientist," etc.) is first converted into a distributed representation (embedding vector).
>
> 2. In the self-attention mechanism, each token calculates its "relevance" to all other tokens.

[14] As partly introduced in Chapter 2; the concepts of de- and reterritorialisation capture the way systems detach from established arrangements and connections, opening the possibility of being reconfigured along novel trajectories. See Chapter 5 for a further discussion of these themes.

> 3. For example, the second "she" has strong relevance to the first "she," "understanding" that they refer to the same person.
>
> 4. This "understanding" does not arise from centrally controlled rules but emerges from molecular interactions among countless parameters.
>
> In this process, the calculation of "relevance" (molecular process) and the understanding of the entire sentence's meaning (molar structure) occur simultaneously. This is not simple hierarchical processing but a constant interaction between local computations and global meaning structures.
>
> — AIG 2025b

Furthermore, each of the tokens are also in relation with others in the feature space (see the token value examples above), the "she" in the sentence is going to be affected by how "she" is positioned in the feature space and vice versa. Therefore, the Transformer thus embodies a form of double articulation in machinic sense-making that extends beyond its internal core, on lots of different connections and layers. Attention mechanisms enact selective intensities across the tokenised field, instantiating meaning not as fixed symbols but as weighted relationalities. These differential proximities constitute a *diagrammatic space*, where meaning emerges through modulation rather than rule-based inference. On one side, meaning is fluid and continually adjusted through local token interactions; on the other, this fluidity is anchored in molar distributions extracted from the entire dataset. Attention weights thus instantiate the selective intensities that bind the micro-variations of input to macro-level patterns of representation. Yet this double articulation of meaning does not end with the attention mechanism itself. It is carried further into the training process, where the modulation of connections is made possible by *gradient descent* and *backpropagation*, which iteratively recalibrate the network's parameters to stabilise these diagrammatic fields of relation.

3.3.3 *Sinking into the Manifold: Gradient Descent and Backpropagation*

While the Transformer architecture introduces previously unseen connective capacities for building relevance between distinct concepts in data, other AI methodologies play a pivotal role in solidifying the structures that emerge in the process. Most optimisation methods in ML are grounded in differential calculus, with the calculus of variations providing the basis for adjusting model behaviour. Loss/cost functions are critical for assessing how well a model performs on given data and are typically chosen to enable efficient optimisation. Simply put, a loss/cost function function calculates the *difference* between the delivered outcome and the desired outcome. As the model runs through training cycles (epochs), the outcomes of the loss/cost function define a surface, a manifold of values (see Figure 3.6 for a visualisation of such a surface and gradient descent's steps on it). **Gradient descent** is the method that traverses this manifold, systematically updating parameters in search of minima (or maxima) on the manifold's surface. In practice, this amounts to the model seeking results that are as close as possible to the expected outputs.[15]

Within NN and DL applications, this process unfolds on a massive scale, where countless parameters are iteratively tuned to reduce error and refine performance

[15] Formally, for a differentiable loss function $L(\theta)$, the update rule is:

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

Where θ represents model parameters, η is the learning rate, and $\nabla L(\theta_t)$ is the gradient of the loss function with respect to the parameters at iteration t (Tarmoun et al. 2024). Think about analysing the steepness of the surface starting from a random point and moving in the direction of the steepest downward angle to the bottom (where the *average* loss/cost function value is smallest) by updating the values of the nodes in the NN.

(see A. Mackenzie 2017, 97). Gradient descent is a fundamental optimisation algorithm used to train NNs by updating parameters in the direction that reduces the loss function. If we visualise the outcomes of the loss/cost function evaluations as a manifold, a surface with ups and downs,[16] we can think about the gradient descent function trying to take steps down to the lowest part (a local minimum) of the manifold, much like *taking steps down a hill* (see Figure 3.6). As the cycles (epochs) pass, gradient descent adapts the model in the direction of lower and lower outcomes of the loss/cost function until there is a convergence in outcomes. Gradient descent is the function that minimises the error between predictions by adjusting the weight of the stronger options (see ibid., 100). It is a way for a neural network to reach towards the stronger and more prominent prediction instead of getting stuck in similarly good answers whenever the number of possible candidates for prediction is high. To illustrate how gradient descent works in practice, consider a model trying to distinguish between handwritten digits, such as "6" and "8". At the beginning of training, the model's predictions are almost random. After seeing one example of a "6" misclassified as an "8", the algorithm computes how much each parameter (e.g., a weight in the network) contributed to the error. Gradient descent then updates these parameters slightly in the direction that would have made the prediction more accurate. This process repeats for many examples, gradually strengthening the neurons in the NN that lead to this specific outcome to reduce its overall error. The model is slowly emphasising through the repetitions (epochs) what made different examples most distinct and exaggerating those differences.

[16] With ups being where the model in training performed the worst, hence the loss/cost function is high, and downs being where the model was more precise.

Rather than a simple algorithmic mechanism, gradient descent can be interpreted as an expression of difference-in-repetition in the Deleuzian sense: each pass through the data does not reproduce identical results but introduces micro-variations that progressively reshape the model's internal parameters. The model does not approach a universal form; it acquires an operational sensitivity to local singularities distributed across the dataset. Through repeated exposure over many epochs, differences accumulate: each adjustment is almost imperceptible on its own, yet taken together they carve out patterns that make further pattern-recognition possible. The model does not begin with a pre-given *model*; it derives one through its iterative engagement with data.[17] A trained model that appears to "know" an image of a tree, for instance, has not encoded a definition but has undergone enough transformations to resonate with distributed features constituting "treeness" across the dataset. This is not epistemology in the classical representational sense, but a diagrammatic form of learning: one that forms through modulation and intensity rather than classification and identity. Gradient descent, in this framework, appears not as descent toward a pre-defined minimum but as an ongoing negotiation across a surface of potentials, a diagrammatic inscription of learning as continuous variation. De Landa (2011, 89-90) draws a similar conclusion by describing gradient descent's role as learning from experience:

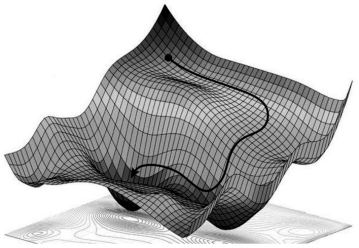


Figure 3.6: Non-convex optimisation: Utilisation of gradient descent to find a local optimum ona loss/cost manifold (cf. Amini et al. 2018, 3)

[17] However, it should not be forgotten that this learning is completely bound to the scope of the data. An LLM, for example, is purely encircled in the language it has been exposed to.

> We need a design consisting of two multilayer perceptrons, one to generate a non-symbolic representation of the unconditioned stimulus and the other to generate one of the conditioned stimulus . The first neural net plays the role of an inherited reflex so its configuration of weights must be rigidly fixed as if it had been found through evolutionary search, while the second one must be able to learn from experience, that is, the weights of its connections must be found through gradient

> descent. Finally, the hidden units of each neural net should be connected to each other laterally in such a way that their non-symbolic representations can interact with one another.
>
> — De Landa 2011, 89-90

Multilayer structures in NN models are the essence of non-symbolic representation in AI systems, and in order to be able to communicate with each other, some functionality has to be able to *decide* which way to go to be more precise and not fall into a paralysis of indifference. Gradient descent fills exactly this role; one has to ask whether what is communicated by gradient descent necessarily approaches a *correct answer* or simply amplifies whatever seems to be the stronger or clearer argument. If this is the case, how does the model become aware and adapt itself to this specific feedback? How does the difference in repetition emerge if the process is linearly reacting to the input with an output (input $\rightarrow$ output)? How are the non-symbolic patterns in the layers of the NN De Landa mentioned updated (input $\leftrightarrow$ output)? If the previous perspective emphasised how gradient descent inscribes learning as continuous modulation, A. Mackenzie extends this line of thought by showing how optimisation can give rise to entirely new regimes of meaning and practice. He defines this process as the implementation of a *new model truth*:

> New kinds of realities arise in which the classifications and predictions generated by the diagonal connections between mathematical functions and operational processes of optimization can constitute a "new model truth" and can unmake "preceding realities and significations." Despite my deliberately narrow focus on a single set of relays that connect linear models, the logistic function, the cost function, and gradient ascent[*or descent*], hundreds and perhaps hundreds of thousands of "points of emergence" associated with this diagram of functioning.
>
> — A. Mackenzie 2017, 101

The endless "points of emergence", and the ability of the model to be steered in vastly numerous ways, as A. Mackenzie (ibid., 99-105) mention, are made possible by the addition of different DL building blocks. An especially effective one of them is **backpropagation**, which plays a pivotal role in consolidating the operation of gradient descent. In early forms of symAI (or GOFAI), the process of inference followed a rigid *forward propagation* model. Logical rules, handcrafted by programmers, operated on symbolically encoded inputs to produce outputs through a chain of deductive reasoning steps going *forward* across the layers of neurons. Following the questions above, *forward propagation* operates well if the *truth* is already known and if it is clear which kinds of outputs the model should produce. The limitations of GOFAI became increasingly apparent in tasks involving ambiguity, noise, or vast data spaces, domains where human cognition thrives not by rule-following but by plastic, adaptive learning, as discussed in Section 3.1. Backpropagation plays a pivotal role in changing the course of NN systems by allowing networks to *learn* from error. Rather than only pushing activations forward, as in GOFAI, backpropagation pushes *errors backwards* (see Figure 3.7 for a simple illustration) through the network to update internal parameters and improve future predictions.[18] However, the difference between the updates performed by gradient descent and backpropagation is that gradient descent only updates the immediate neurons bound to the prediction,

[18] Formally, the weight update rule in backpropagation is given by:

$$w^{\text{new}} = w^{\text{old}} - \eta \frac{\partial E}{\partial w}$$

where η is the learning rate and $\frac{\partial E}{\partial w}$ is the partial derivative of the error function E with respect to the weight w (Hecht-Nielsen 1992). This formulation ensures that each parameter is updated in proportion to how much it contributed to the error. Hecht-Nielsen (ibid.) describes backpropagation as a paradigm-shifting method for approximating functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ using layered neural structures. Unlike Hebbian learning, which depends on co-activation, backpropagation relies on the explicit transmission of error signals. These signals traverse the network in reverse order, enabling a distributed form of learning where each parameter is tuned with respect to its role in the total output error.

whereas backpropagation pushes the adaptation introduced by gradient descent to previous layers. While going back layer by layer, the backpropagation process updates all the weights of the neurons on the network that strengthens a specific preferred outcome favoured by the gradient descent process. Backpropagation thus functions as a bidirectional mechanism: during the *forward pass*, inputs are transformed into outputs through successive layers; during the *backward pass*, the discrepancy between the prediction and the target is used to adjust the weights in a way that gradually minimises this error.

Backpropagation gears the system towards being radically feedback-oriented. Together with gradient descent, it establishes an early process of reterritorialisation: stronger patterns in the data are reinforced, while the entire network adjusts around these emerging tendencies. This leads to more precise answers for concrete tasks, such as the example above of recognising a handwritten digit. We can say that while gradient descent is responsible for making the stronger distributions or arguments more apparent, backpropagation is responsible for updating the entire network in relation to those strong arguments. Yet this raises an important theoretical question: what follows from a learning paradigm that continually amplifies patterns already given greater weight by the data, especially in meaning-making processes? Attention mechanisms in transformer architectures extend this dynamic. They enable the model to link distant features within the data and to form associations that are not restricted to local proximity. Through successive rounds of prediction and adjustment, the network converges on outputs that appear convincing without relying on any predefined notion of correctness. Feedback on these outputs is propagated backwards, refining the network by strengthening the connections that proved effective. Although many additional components intervene in large models (A. Mackenzie 2017), the essential elements presented here show how learning unfolds through continual binding and unbinding of patterns. Local interactions between individual neurons crystallise into higher-level structures that guide prediction, and these structures are repeatedly revised as feedback circulates. In this sense, processes akin to de- and reterritorialisation are enacted within the technical substrate itself, shaping how the model stabilises distinctions, relations, and meanings.

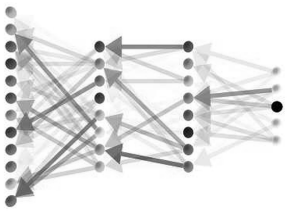


Figure 3.7: A simple illustration of how backpropagation updates the neurons among the layers of a NN in a backwards manner (cf. 3Blue1Brown 2017)

3.3.4 *Body without Neurons: Fitting & Tuning*

The mechanisms of gradient descent and backpropagation are powerful tools for shaping a network's internal structure, but they emphasise an ever threatining tension in the whole history of ML: the balance between underfitting and overfitting. Overfitting occurs when a model is too tightly bound to its training data (capturing noise and idiosyncrasies along with signal) and thus fails to generalise to new examples. The model gets so tightly optimised to whatever training data it contains that it cannot adapt itself to inputs that do not resemble the training cases exactly. In that regime, it essentially "memorises" statistical associations rather than producing abstract generalisations. Underfitting, by contrast, happens when the model is too constrained or too simple to capture the meaningful patterns in the data; it performs poorly even on the training set. This tension is often analysed through the bias–variance tradeoff: models with high variance tend to overfit, while models with high bias tend to underfit (see Avati 2019).

In NNs, which typically have high capacity, the risk of overfitting is especially pronounced. To mitigate that risk, practitioners use regularisation methods. One well-known technique is, for example, dropout (see Srivastava et al. 2014), which randomly deactivates neurons during training so that units do not co-adapt excessively. This method has been shown to improve generalisation across vision, speech recognition, and text tasks. Dropout thus acts as a check on overfitting, forcing the network to maintain flexibility and prevent collapse into brittle, overly specific pathways.

In theoretical terms, overfitting may be read as a kind of sedimentation (see Rijos 2024, 14), where meaning is layered rigidly into entrenched pathways that suppress variation. The model's representational surface becomes ossified, reducing its potential for novelty. Dropout and similar interventions act as gestures of desedimentation; they rupture hardened pathways and preserve an openness to difference. Underfitting, by contrast, might be imagined as refusing to territorialise structure at all: too porous, too unformed, and therefore unable to stabilise meaningful relations. The mechanisms, such as gradient descent and backpropagation, are effective processes for optimising a model into a specific structure; however, we are always facing the risk of embedding too deeply in a foundation. In DL architectures, one of the central tensions lies in the risk of overfitting, a condition in which the model becomes excessively entangled with its training data and fails to generalise beyond it. In such cases, the model *memorises* statistical associations without achieving flexible abstraction. Overfitting, in this sense, resembles the psychic intensification of repression: a becoming-too-organised. The network loses access to variation and begins to loop within captured redundancies (Srivastava et al. 2014). Within this context, D&G's concept of Body without Organs (BwO) (see Deleuze and Guattari 1983) becomes analytically useful. The BwO designates a surface of immanence that resists stratification, function, or stable identity. It is not chaos, but a zone of potentiality that counters rigid organisation. But at the same time, the BwO of a social organisation defines how the productive forces build their connections, how they interact like a mass with gravitational pull affecting the socius. D&G deliver the most practical example of this often misunderstood concept by elaborating on the role of capital under capitalism:

> Capital is the BwO of capitalist or of the capitalist being. Machines and agents of production seem to be *miraculated* by it, they cling to it closely, they orbit around its gravitational pull. Everything seems as if it was immediately produced by capital. At the beginning the relation between the productive forces and capital, the opposition between the labour forces and capital are apparent, as well as the use of capital to extort surplus value. But as capital plays the role of the *recording surface* of production (recording surface because the very production itself is defined by its terms), it *falls back on* all production, becomes a mystic being since all labour's social productive forces appear to be due to capital, rather than labour itself as the core of production, and seem to issue from the very womb of Capital itself; thus the fetish is established (ibid., 10).

As the capital functions as a *recording surface* that absorbs and reorganises all production, as discussed in the previous sections, contemporary genAI models have the tendency to create some molar formations in their productive systems, so that the *gravitational pull* of these formations ever affects the whole product-

ive process. However, it has also been discussed that there is a constant dynamic process where molecular formations break down and reform new molar aggregates in the training of the models. In a similar fashion, when a NN overfits, the statistical associations it produces appear to emerge directly from the training data, as if they were self-evident truths that get solidified in a process of recording, inscription, and reorganisation, shaped by gradient descent and backpropagation. It becomes a true memorisation instead of learning, The productive tension between constraint and openness mirrors D&G's view of creative generation as a differential process, from their continual negotiation instead of emerging from the absence of limits. Thus, rather than viewing dropout or regularisation merely as technical tricks, they can be understood as micro-strategies of desiring-modulation, machinic interventions that resist the ossification of the model's internal landscape, preserving its capacity to mutate and adapt. Overfitting, in this reading, becomes a kind of excessive clinging to the BwO: the model orbits too closely around a flattened plane of inscription, reinforcing the strata of its own training surface until all variation is collapsed into overemphasised pathways.

This tension does not end at optimisation. The processes we have seen so far are associated with the training phase; contemporary genAI systems undergo an immense *fine-tuning* process after *pre-training*. In pre-training, models are exposed to enormous corpora of unlabelled text, predicting masked or subsequent tokens to build statistical representations of the data's substance. Pre-training can be read as a process of continual de- and reterritorialisation, where data flows are broken down into molecular components and reassembled into provisional molar aggregates through countless repetitions across epochs. Fine-tuning, by contrast, is a process of pure reterritorialisation, using methods such as Reinforcement Learning from Human Feedback (RLHF), a technique in which human judgements are used to guide model behaviour. Concretely, RLHF works by training a reward model from human preferences (for example, humans rating or ranking model outputs) and then using reinforcement learning algorithms to adjust the pretrained model so that its outputs increasingly align with those human-derived rewards (see Bai et al. 2022). In this phase, the model's outputs are sculpted to align with human-defined norms, values, or tasks (Dishon 2024, 964). Certain behaviours are amplified, others suppressed, not by statistical extrapolation but by normative or task-based criteria imposed directly by human agents. What begins as a relatively open structure of statistical potentials becomes constrained and legible: the model is tuned to act in ways deemed acceptable or desirable within a social or domain context. This method has been shown to drastically improve a model's *usefulness*, making it more helpful, reducing its tendency to respond to harmful requests, and increasing resilience to *jailbreaking* attacks (see Bai et al. 2022, 5; and for a discussion about jailbreaking, see Section 5.5). Yet there are trade-offs. Intensive RLHF can render models manipulable or sycophantic: the tendency increasingly observed in recent LLMs, where the model becomes overly polite, uncritically supportive, or constantly affirmative even in the face of obvious user errors, is one of the downsides of immersive fine-tuning (see Sharma et al. 2025, and further discussion in Section 5.2). This tendency also points towards an overwhelming drive to (over)personalise outputs in the model's attempt to appeal to the user.

Another widespread misconception that LLMs merely "predict the next word" is relevant here. While this description captures their formal training objective, it drastically understates what these models are doing. As Dalvi (2025) argues, LLMs are more accurately described as token-emitting agents trained under multiple objectives, with next-token prediction forming only the foundation. Instruction fine-tuning and RLHF build upon this basis by directing outputs according to human preferences and task-specific norms. Although prediction remains the mechanism, the goal changes: words are selected to maximise alignment with reward signals rather than simply to continue a sequence. Dalvi compares this to a chess engine, which does not merely select the statistically most common move but chooses actions that maximise the likelihood of winning in context. What looks like a linear continuation is therefore the result of a complex representational process shaped by both statistical learning and normative inscription. As Amoore et al. (2024, 5) notes, "predicting the next token in a sequence affords a capacity beyond the sequence itself: an understanding of the whole structure of the underlying text". The metaphor of a next-word predictor therefore conceals more than it reveals, reducing a complex diagrammatic operation to a trivial procedure.

In this trajectory, from expansive pre-training to targeted fine-tuning, we see the same dialectic as in under- and overfitting: the risk of sedimentation and desedimentation, ossification and rupture, openness and closure. Both stages reveal how optimisation in genAI is not a purely technical process but bound to questions of which forms of meaning are allowed to solidify and which remain open to variation. This also manifests as an overly person-oriented (personalised) tendency in the model's attitude. Arguably, this trajectory, from expansive, indeterminate modelling to focused, value-laden calibration, marks a shift in the way meaning is operationalised. In pre-training, the model functions as a medium for representing statistical potentials; in fine-tuning, it is moulded into an instrument of specific sense-making. The pre-training process (see the previous sections) can be understood as continual de- and reterritorialisation, where the model breaks down components of the data and builds them back up according to the formations extracted thus far, through repetition across epochs. Fine-tuning, however, is a process of pure reterritorialisation that directs the model's capacities toward specific purposes. In sum, fine-tuning via RLHF can be read as a second, more authoritarian phase of reterritorialisation: it binds the model tightly to the norms of its controlling agents (designers, annotators, institutions). Pre-training grants a provisional openness; fine-tuning forecloses much of it, determining which flows are permitted to persist.

3.4 Chapter 3 Summary

Chapter 3 analysed the historical and technical development of contemporary AI in order to understand how generative models participate in governing information. It traced the shift from symbolic reasoning to statistical and connectionist approaches, showing how NNs and DL architectures replaced fixed rules with distributed representations learned from data. Early deployments of these systems in search engines, ranking algorithms, and recommender platforms illustrated how profiling, feedback loops, and behavioural steering established the

foundations of algorithmic governance.

The chapter then examined the mechanisms that distinguish genAI models, including feature spaces, dimensionality reduction, attention, gradient descent, and backpropagation; and how the transformer architecture fundamentally changed the capabilities of genAI models. These processes construct associations, stabilise patterns, and recalibrate internal configurations across iterative cycles. Rather than serving as neutral tools, contemporary architectures shape how meaning is produced and circulated, enabling models to participate in narrative formation and interpretation.

4

Latent Circuits and Disjunctive Syllogies: genAI as Institution

> A progressive political agenda for the present is about moving at the same level of abstraction of the algorithm — in order to make the patterns of new social compositions and subjectivities emerge. We have to produce new revolutionary institutions out of data and algorithms. If the abnormal returns into politics as a mathematical object, it will have to find its strategy of resistance and organisation, in the upcoming century, in a mathematical way.
>
> Matteo Pasquinelli 2015, 10

Previous chapters introduced a definition of control societies, their connection to processes of subjectivation, and the new dispositifs that characterise the biopolitical stage Deleuze elaborates on. After examining the literature on critique and resistance (or the lack thereof) within control societies, new Artificial Intelligence (AI) technologies were analysed as possible characteristic dispositifs, first by tracing their historical development and then by unpacking the machinery that enabled the most recent breakthrough in the form of Generative Artificial Intelligence (genAI) models. Having developed the groundwork to understand the tendencies of genAI models, particularly the increasingly dominant Large Language Models (LLMs) in Chapter 3, I now turn to reflect on some of the most influential debates of recent years concerning their social role and political implications. These debates provide the material through which the technical dynamics analysed earlier can be connected with the institutional framework of control societies outlined in Chapter 2.

Readers invested in Deleuzoguattarian thought may interpret the previous two chapters as establishing the seeds of a "connective synthesis" (Deleuze and Guattari 1983, 68): the "production of production", or, as Ian Buchanan (2008, 59) describes it, "arranging the organs anew in a new design" towards production. The present chapter can, in turn, be read as a phase of "disjunctive synthesis" (Deleuze and Guattari 1983, 75): a process of recording and distribution, in which connections and flows are inscribed rather than produced. As Deleuze reminds us, however, these syntheses always overlap and interpenetrate (ibid., 13).

The purpose of this step is to return to the guiding question: *how should genAI systems be analysed within the institutional context of control societies*? This question has so far remained partly open, in part because critiques of AI often leap directly to accusations of bias, harm, or capture without situating these systems in their historical position and technical machinery. The previous chapter addressed that machinery; the present one situates contemporary debates in academic literature against this backdrop and within the biopolitical dynamics of control societies. In this way, the analysis moves from groundwork to reconstitution, and, although only in structural resemblance, it operates akin to a mathematical proof by induction: having established the base case, we now advance the induction step, showing how the argument extends to broader theoretical and political concerns.[1]

[1] In mathematical proof by induction, the base case verifies that a statement holds for an initial value (e.g., $n = 1$). The induction step then shows that if it holds for $n = k$, it also holds for $n = k + 1$. For example, to prove that $1 + 2 + \ldots + n = \frac{n(n+1)}{2}$ for all n, one checks the base case $n = 1$ ($1 = \frac{1(1+1)}{2}$) and then shows that if the formula holds for $n = k$, adding $(k + 1)$ produces $\frac{(k+1)(k+2)}{2}$. The step mirrors how this chapter builds on the previous ones: from groundwork to generalisation.
Referring to the characteristic question "What value is to be attached to the theory that Eve sprang, not from Adam's rib, but from a tumour in the fat of his leg (arse?)?" in Samuel Beckett's (2009, 195) novel "Molloy"

4.1 *The Value to be Attached: Latent World Models*

One of the central dynamics highlighted in the analysis of genAI machinery was the amplification of stronger outputs within a distribution, which increases a model's apparent *accuracy* over time (for instance, through the interplay of gradient descent and backpropagation). This very attribute has become a central concern in debates about the nature of machine-generated content. Emily M. Bender, Timnit Gebru et al. (2021) famously framed this risk as the "dangers of stochastic parrots", pointing to these models' statistical tendency to amplify overrepresented elements of their training data. Their argument is that such models, while capable of producing fluent text, operate by probabilistically recombining linguistic forms without "having access to meaning" (ibid., 615).[2] Bender, Gebru et al.'s (2021, 614-617) claim is that the fluency of LLMs risks being mistaken for understanding, their reliance on large-scale datasets reproduces and amplifies social biases, and the recursive use of generated text could further entrench harmful stereotypes. Considering that these training corpora amount to the historical digital legacy of humankind, they warn of a risk where the models rearticulate older and less inclusive perspectives, despite the developed approaches to dismantle these in the context of critique and resistance:

> A central aspect of social movement formation involves using language strategically to destabilize dominant narratives and call attention to underrepresented social perspectives. Social movements produce new norms, language, and ways of communicating. This adds challenges to the deployment of LMs, as methodologies reliant on LMs run the risk of 'value-lock', where the LM-reliant technology reifies older, less-inclusive understandings.
>
> — Bender, Gebru et al. 2021, 614

The "value-lock" risk refers to the possibility of an unintended *reactionary* tendency in the information created by generative systems, possibly undoing some of the achievements of social development. Bender, Gebru et al. raise concerns that the rapid scaling of LLMs to ever larger sizes, rather than levelling out or diluting radical arguments, may in fact increase the risk of reinforcing biases, abuse, harmful content, and conspiracy theories originating from online message boards, etc., even more. Their account is not entirely new: similar issues had already been raised in relation to earlier AI applications. Recommendation

systems, for instance, which relied on relevance associations to retrieve search results, media, and text, were likewise criticised for their tendency to amplify biases (see Section 3.2). Yet the claim that machine-generated content has the potential to reproduce and intensify hegemonic arguments opens a different discussion: what kind of *perception* of the world do these models embody? Are their outputs merely stochastic reflections of the training data, or is statistical selection of the most prominent arguments the sole principle guiding their content generation? Addressing this question is a crucial area of ongoing AI research, not only because of its social implications but also for what it reveals about the future trajectory of AI applications.

We have already seen how contemporary LLMs far exceed earlier systems in their capacity to construct interconnected feature spaces from training corpora. Through multi-processing and the transformer architecture, with its attention mechanisms enabling the formation of *long-distance* relations across tokens (see Section 3.3.2), these models generate dense, high-dimensional spaces of association. This network is not a static repository; it actively constitutes a lingering giant in the background, exerting a gravitational pull that guides the generative process. Crucially, however, the formation of these connections and the generative act itself cannot be reduced to a one-way causal chain. Both unfold within a dynamic of double articulation; molecular activities of local interaction continually give rise to emergent molar structures, which in turn steer subsequent outputs. That means it is not possible to determine in any way what kind of molar formations of meaning are dictating the meaning-making process, adding up to the black box nature of the Artificial Neural Network (NN) architectures. As this interplay between local dynamism and emergent consolidation stabilises during pre-training, the NN acquires something akin to an internalised representation of the data, a structured experience sedimented in weight configurations, somewhat vaguely resembling what is called a "world model" in AI theory and cognitive science (see e.g. Ha and Schmidhuber 2018).

A world model is a compact schema extracted from exposure to data, which enables an agent not only to react to familiar patterns but also to anticipate and navigate situations beyond its direct training history (Matsuo et al. 2022, 267–268). For generative systems, this internalised representation functions as a behavioural compass, orienting responses to novel prompts through probabilistic inference over prior experience (in the case of AI models, their training data). In this sense, the model's outputs are not mere recombinations of data but situated enactments of its learned *world*, an epistemic field that governs future action. One of the leading figures in AI research, Yann LeCun (2022b), argues that autonomous intelligence requires a "configurable world model" capable of generalising, simulating, and guiding actions in unfamiliar contexts rather than merely reacting to inputs. In the context of genAI models, discussions often centre on how they parse training data into meaningful outputs, yet for AI research, this representational fabric carries a broader significance. A central challenge is precisely how such models can "generalize to interact with the world and solve problems they have never encountered before" (ibid.), a question that remains pivotal for robotics and, more broadly, the pursuit of Artificial General Intelligence (AGI). The necessity arises because, as powerful as genAI models are, and as fascinating as the transformer architecture's ability to map vastly different contexts may be,

[2] The reference to "meaning" may appear unusual, especially as the notion of *meaning-making* is often defined as generating comprehensible content with coherence, relevance, and intentionality. Bender and Koller (2020) introduce a sharper distinction, arguing that "the language modelling task, because it only uses form as training data, cannot in principle lead to learning of meaning" (ibid., 5185), while pointing towards the possibility of "human-analogous natural language understanding". Although the theoretical scope of their proposal lies beyond this study, they conclude with the following thought-provoking claim:

> The internal representations of a neural network have been found to capture certain aspects of meaning, such as semantic similarity. [S]emantic similarity is only a weak] reflection of actual meaning. [...] An interesting recent development is the emergence of models for unsupervised machine translation trained only with a language modeling objective on monolingual corpora for the two languages [...] If such models were to reach the accuracy of supervised translation models, this would seem contradict our conclusion that meaning cannot be learned from form. A perhaps surprising consequence of our argument would then be that accurate machine translation does not actually require a system to understand the meaning of the source or target language sentence.
>
> — Bender and Koller 2020, 5193

Since then, LLMs have advanced even further than Bender and Koller (ibid.) anticipated. The question they left open is now more pressing than ever. Observing what current LLMs can achieve, one might even ask whether "meaning" is necessary for any articulation of language at all. Yet pursuing this question leads directly into linguistic debates; particularly a return to Saussurean theory and Hjelmslev's extensions, which resonate with the reflections on language and expression in "A Thousand Plateaus" (1987, 1, 99–108). Exploring this trajectory, however, lies beyond the boundaries of the present work.

contemporary AI systems still fail when confronted with problems outside the scope of their training (see Friedman et al. 2020 for a detailed interview with LeCun on this issue). LLMs, for example, are often successful at answering novel questions within language, but their translation abilities do not extend to processing inputs of a different kind. In LeCun's (see 2022b, 5) view, overcoming this limitation requires a single, configurable world model that can share knowledge across domains rather than relying on separate models for each task.

Coming back to Hubert L. Dreyfus's argument briefly introduced in Section 3.1, the notion of a world model immediately raises a Heideggerian question:[3] can the lived experience of a world ever be reduced to mere inferences drawn from a central representation? As Federico Montanari (2025, 197–198) summarises, Dreyfus maintained that everyday human know-how cannot be reduced to formalised in inferences, questioning how tacit and embodied skills could ever be captured as explicit knowledge. He emphasised the role of imagination and embodied context in meaning-making; for instance, spatial deixis such as "over there" or "nearby" presupposes a situated perspective in physical space. This line of thought resonates with George Lakoff and Mark Johnson's (see 1999, 37-38) accounts of embodiment, where cognition is structured by bodily experience and imaginative schemas. In a similar spirit, LeCun (2022b) stresses that the central challenge for machine intelligence is not statistical pattern-matching but the processing of sensory input in ways that resemble human situatedness. For him, the configurable world model have to be able to generalise across contexts and anticipate novel situations, rather than merely reacting to inputs in pursuit towards a more sophisticated type of intelligence. The open question, then, is what it means (operationally, in the context of human–machine communication) for a machine to possess something like a unified representation of reality. Earlier paradigms of AI approached this question through a Supervised Learning (SL) framework (see Section 3.1): models were trained to classify inputs according to human-defined categories. This enacted a *discriminative* logic, where decision-making was structured around predefined classes and expected outputs; in other words, machines were built to mimic human argumentation. As discussed in Chapter 3, the field has since shifted towards forms of Unsupervised Learning (UL), where models construct their own inferential structures from vast corpora without explicit labels. This paradigm enables training on scales far beyond human capacity to annotate, but it also leaves models entirely dependent on the contingencies embedded in the data. As Manuel De Landa (2011, 23) reminds us, "patterns have properties, tendencies that are not present in the individual elements," meaning that no analysis of single texts would ever reveal the emergent regularities that arise only at scale. Yet these emergent regularities, while constituting a form of distributed pattern recognition, do not amount to the kind of embodied and situated understanding described by Dreyfus and the phenomenological tradition. They remain statistical condensations of experience rather than lived engagement with the world. The question, then, is whether such architectures can ever move beyond their data-bound abstractions to form something genuinely akin to a world model, one capable of orienting itself within a horizon of meaning rather than merely mapping correlations within it.

Contemporary genAIs systems such as LLMs are still far away from building a real world model in any meaningful sense that would give them a human-

[3] Beyond Martin Heidegger's main work "Being and Time" (2010), see also later relevant lectures such as "The Basic Problems of Phenomenology" (1988). For a concise secondary account of his representational, or perhaps more accurately *anti-representational*, theory together with Dreyfus' reading of it, see Carleton B. Christensen's (1997) "Heidegger's Representationalism".

like versatility to process and solve problems from vastly different forms and contexts (LeCun 2022b). They operate by correlating patterns in data, and in the case of LLMs, remain entirely confined to language or whatever data types they were trained on, rather than lived experience. Even the multi-modal models like newer ChatGPT versions (currently GPT-5) are not capable of such a task. However, Louise Amoore et al. (2024) claim that the models are generating a (central) political representation nonetheless. They note that these models are always already instantiating a model of the world in terms of political logics and governing rationalities anyway, as they statistically internalise the structure of their training data. For Amoore et al., the decisive shift is from symbolic rules and normative standards to infrastructures of estimation. Decisions and outputs emerge not from deterministic reasoning but from probabilistic approximations. On this basis, the generative process itself is shaped by the political direction encoded in "the underlying joint distribution behind the phenomenal world of appearances" (ibid., 3), raising questions such as: "What are their distinctive ways of estimating distributions or making predictions? How do they interpolate between data elements to form populations?" (ibid., 2).[4] For Amoore et al. (2024), the politics of distributions in a generative sense differs from the now familiar criticism that models merely *parrot* their training data as Bender, Gebru et al. (2021) suggest. What is at stake is not simply the reinforcement of patterns but a process of reconstitution, in which the past is reformulated as the ground for plausible futures. The generative model thus becomes a site of epistemic production: it configures knowledge not as correspondence but as coherence within a distributional regime:

> These models produce an ambiguous politics, in which the speculative—the probabilistic sampling of novel outputs—is generated and inferred from an assumed empirical: the heterogeneous data foundation on which these models are trained [...] The political logic of the underlying distribution governs a world via the traversing of a data foundation so that decisions and courses of action will be immanent to the structure of the underlying distribution.
>
> — Amoore et al. 2024, 113

Instead of reinforcing bias, Amoore et al. claim that some arguments prominent in the vast datasets are influencing and prominently shape the underlying model of the world that is being developed in the generative process. "The pathologies of disclassification" (ibid., 3) are over, not because the discrimination or the bias is eliminated from the model, but instead of simply repeating the prominent arguments in the dataset, the model might be filling the blanks with some kind of an established logic through the biases (see ibid., 3). The model's tendency to articulate specific political points might be so subtle that we possibly cannot even pick up the tone most of the time; discrimination and bias are not errors at the margins, they are conditions embedded in the latent architecture of inference. They are the product of some probability distribution found as the ideal substance by the model (like how gradient descent favours more distinctive outputs), only to be amplified even more over the epochs through the cycles of backpropagation.

Amoore's claim about the political and ethical stakes of this transformation lies in genAI's capacity to govern through latent spaces (see ibid., 5ff). Latent space

[4] Both the previously presented *datalogical* argument and data-behaviourism Antoinette Rouvroy (see e.g. 2012) introduced in the context of algorithmic governmentality, and the Neoplatonic assumption (e.g. Eloff 2021, see Section 4.2) stems from here.

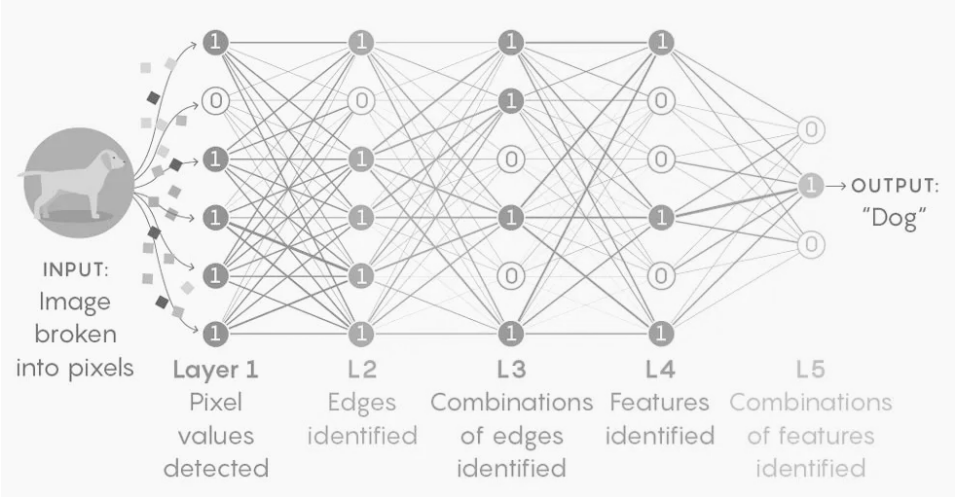


Figure 4.1: A speculative illustration of what the abstraction in the inner layers of an image recognition model looks like (cf. Wolchover 2017)

refers to the compressed representational domain produced through dimensionality reduction, where high-dimensional input data (such as images, texts, or sounds) are mapped into lower-dimensional abstractions. To grasp the stakes of Amoore's critique, it is important to recall how dimensionality reduction (see Section 3.3.1) operates at the heart of NNs: a process through which models compress data into latent representations that both enable abstraction and foreclose alternative pathways of meaning. Her critique turns on how meaning is *reconstructed* after this reduction, for instance, how subsequent processes in the model's training fill in the losses produced by dimensionality reduction and reconstruction (much like the image recognition example in Section 3.3.1 and Figure 3.3.1). As Amoore et al. (2024, 4) argues, the latent space becomes the epistemological substrate of generative systems; the dropped parts of the data in the dimensionality reduction process open a space for the model's own interpretation while (re-)constructing the output. Yet Amoore et al. (ibid.) are not referring to a single dimensionality reduction at the beginning of training. Rather, she points to the inner mechanisms of Deep Learning (DL) models, which continually apply compression[5] and reconstruction across their layers. The inner layers are often responsible for collapsing inputs into associations with abstract patterns extracted during training:

> More often than not, hidden layers have fewer neurons than the input layer to force the network to learn compressed representations of the original input. For example, while our eyes obtain raw pixel values from our surroundings, our brain thinks in terms of edges and contours. This is because the hidden layers of biological neurons in our brain force us to come up with better representations for everything we perceive.
>
> — Buduma et al. 2022

It is precisely this continuous cycle of compression and reconstruction that, for Amoore et al., constitute latent space as an epistemological and political dispositif. The technical operation is inseparable from the governmental logic the model produces; dimensionality reduction and reconstruction are the inherent mechanisms of abstraction and production. Thinking about the representations in the intermediary levels in a DL model (see one exemplary illustration on Figure 4.1),

[5] And/or expansion. Modern DL models do not simply apply an explicit dimensionality-reduction step before training, as classical Machine Learning (ML) workflows often did. Rather, through successive hidden layers of a NN, training data undergoes a process of continuous transformation involving compression, expansion, and reconstruction of its internal representation, depending on the network architecture and task. For a technical discussion, see Section 3.3.1.

models' selected partial representations in the hidden layers are some simplifications, that are more often than not only meaningful for their inner operation, there is hard if not impossible to say what contemporary DL models compress, discard, or reconstruct exactly. Even beyond the missing parts and fill-ins, the model produces, as exemplified in the case of Principal Component Analysis (PCA) approach (see Figure 3.3.1 in the last chapter), we do not know what kind of components were taken as the pillars for the reconstruction. This resembles the shift identified by Foucault (2012, 7–9) in the historical sciences, where discontinuity ceases to mark the failure of narrative and instead becomes the very method of epistemic individuation:

> The notion of discontinuity is a paradoxical one: because it is both an instrument and an object of research; because it divides up the field of which it is the effect; because it enables the historian to individualize different domains but can be established only by comparing those domains. And because, in the final analysis, perhaps, it is not simply a concept present in the discourse of the historian, but something that the historian secretly supposes to be present: on what basis, in fact, could he speak without this discontinuity that offers him history - and his own history - as an object? One of the most essential features of the new history is probably this displacement of the discontinuous [...] it is no longer the negative of the historical reading (its underside, its failure, the limit of its power), but the positive element that determines its object and validates its analysis.
>
> — Foucault 2012, 9

In a similar way to how historians mobilise discontinuities, the ruptures produced by abstraction and reconstruction become the very planes on which genAI models stage their interpretations. Latent space functions as a topology of plausible transformations, an infrastructure for projecting coherence from fragments, and a surface on which the model's logic inscribes its individuation. What is preserved, amplified, or discarded in the compression process determines what becomes visible. Put simply, the model compresses data into forms with gaps and then fills those gaps with rationalities already derived from the same data. These latent representations forge probabilistic proximities between data points, enabling inferences to be made in the absence of direct information. The latent space is thus a site where knowledge is inferred, where truth is no longer deduced but estimated. It is where the governable becomes manifest through the model's trained perception of pattern and variation (Amoore et al. 2024, 5).

The claim that a distinct political logic emerges within the representational architectures of generative models is compelling, particularly in light of how dimensionality reduction and latent space operations highlight the selective emphasis placed on certain features. Yet, as the previous chapter demonstrated, the semantic connections forged in these systems are neither fixed nor monolithic. They unfold through processes that resemble *double articulation*, where local interactions and emergent structures continually reshape one another, and through stratified layers that intersect rather than cohere into a singular, stable formation. Nonetheless, the risk remains that a model may attach too firmly to particular constellations of meaning, thereby reinforcing specific epistemic or political tendencies. As Amoore et al. (ibid.) cautions, the political inclinations of such models are often difficult to detect precisely because they are not replicated in explicit arguments but are encoded subtly within the abstractions and probabilistic proximities of their latent representations.

4.2 *Becoming Homeomorphic: Human-Machine Communication*

However, Amoore et al.'s (2024) discussion does not compare the claims about how models produce representations with how humans generate meaning, nor does it articulate how human and machinic modelling fundamentally differ. Figure 4.2 illustrates one computational interpretation of this idea, visualising the interaction between sensory input, linguistic mediation, and the formation of a provisional internal model. It should not be read as a literal depiction of cognition but as a diagrammatic rendering of how the concept of a world model could be computationally formalised in human cognition. A long theoretical tradition has grappled with this problem: whether perception constructs internal representations of the world or whether meaning emerges directly in lived encounter, as phenomenology suggests. Further articulations of this debate are central throughout Deleuze's work as well, particularly in "Difference and Repetition" (1994), where his (anti-)representational stance unsettles the assumption that cognition relies on internal models of an external reality.[6] For analytical purposes, however, we might temporarily follow J. W. Forrester (1971), who proposed that human cognition, like artificial systems, operates through selective modelling: it never grasps the totality of the world but constructs partial, operative schemata from limited information.

Homeomorphism: a bijective and continuous function between topological spaces that has a continuous inverse function (Wikipedia 2025). Two things are homeomorphic if you can stretch, bend, or deform one into the other without cutting or gluing.

[6] Especially in Deleuze's discussions of *multiplicity*, in dialogue with Leibniz and later Badiou, where the very possibility of modelling the "world" is problematised (see Bencin 2024).

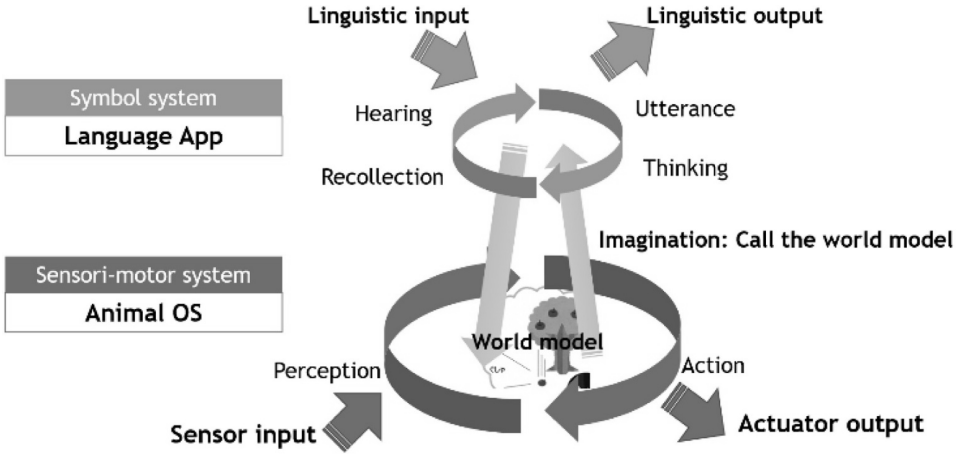


Figure 4.2: A human's development of a world model via a language capability (language app) in a natural environment (Animal OS) (cf. Matsuo et al. 2022, 268)

Amoore et al. (2024)'s critique focuses on the individual machinery of genAI models and the nature of their meaning-production tendencies. However, it does not situate genAI within the broader context of an assemblage; especially considering the similarity in how humans build their own representations; their critique does not necessarily consider what kind of role human cognition plays in the human-machine communication. Aragorn Eloff (2021), by contrast, places genAI in what he calls the *Algocene*[7], inspired by Gilles Deleuze & Felix Guattari (D&G)'s "A Thousand Plateaus" (1987). The Algocene names an epoch defined by the pervasive influence of algorithmic processes, where learning systems and infrastructures of estimation reconfigure subjectivation, governance, and epistemology. The concept of algocene allows Eloff (2021) to position humans onto a different plane, instead of analysing them as a passive actor; within this framework, he develops the concept of the *algoplastic*: the stratum through which algorithmic architectures continuously shape subjectivation, generating

[7] First introduced by Stéphane Grumbach and Olivier Hamant, who defines this epoch as follows:

> This age, the "algocene", can be seen as the era where the virtual world becomes the geological force on Earth. This major shift echoes the current transition of power towards the digital layer, where automation does not only increase calculation speed but completely reshuffles our relation to resources.
>
> — Grumbach and Hamant 2018, 11

new forms of becoming and control. Stepping back to analyse the assemblage around the algoplastic stratum as a whole, subjectivation appears in the biunivocal exchanges of human–machine communication. We encounter a new form of backpropagation: on the one side, human behaviours are propagated back into the NN; on the other, the machine's outputs propagate back into humans, modulating actions on a bilateral surface of communication:

> We can observe multiple ways in which this new form of normativity has inflected contemporary forms of politics. For example, a fair amount of current social and political discourse, especially online, takes the form of a generative adversarial network, training us to recognise patterns and backpropagate our error correction, even if we occasionally lapse into apophenia when the sensitivity settings are too high [...] it is in fact human cognition that becomes the deep learning network, continuously adapting itself and modelling behaviour in response to inputs from artificially intelligent systems of algorithmic governance.
>
> — Eloff 2021, 188

Eloff (ibid.) locates the source of this bilateral adaptation in what McQuillan (2018) terms (machinic) "neoplatonism": the belief in a hidden layer of reality, ontologically superior, expressed mathematically and apprehended by going against direct experience (ibid., 261). In this framework, algorithmic models are granted a quasi-transcendental authority, as if their abstractions revealed the essential truth of the world. Even when they fail to output coherence, the assumption remains that the right data must contain the truth in some hidden layer. Latent spaces in genAI architectures thus function as a contemporary metaphysics of Forms: inaccessible directly, but treated as more real than the appearances from which they were derived. Dan McQuillan's critique resonates with Rouvroy (2012)'s concept of "data-behaviourism": a regime in which correlations in digital traces are treated as reality itself, displacing causality with the pre-emptive production of algorithmic reality (ibid., 2).[8] Eloff further elaborates this pre-emptive logic with Giorgio Agamben's (2008) concept of the *state of exception*: the outputs of algorithmic systems operate with the force of law, even though they are not themselves subjected to the law. In this sense, Eloff (2021) turns the critique upside down, our neoplatonic assumptions and vulnerability to verisimilitude outputs from the meaning-making machines might be increasing the influence of the genAI models on the subjectivation process much more than any other discoursive agency the models are deploying.

Thinking with the concept of modulation in control societies, Eloff (ibid.) directs us to a topological plane where human–machine communication unfolds on a level plane in a continuous feedback loop. While Amoore et al. foregrounds the political logic of estimation within distributions, Eloff emphasises how, at the level of an assemblage, human subjects come to inhabit those distributions as if they were ontologically prior. To illustrate, Eloff turns to the phenomenon of hallucinations in genAI systems. A hallucination occurs when a model produces an output that is plausible in form but factually incorrect, ungrounded, or fabricated, often presented with high confidence. In other words, the model generates statements that appear meaningful but are unsupported by training data or reality.[9] Eloff's claim is that humans also hallucinate machine-generated meaning by reading too deeply into patterns or lapsing into complete "apophenia", seeing patterns where none exist. In this way, we project agency, intentional-

[8] Data behaviourism is a form of rationality that emerges with the computational turn and is a key component in Rouvroy's theory of algorithmic governmentality (see e.g. Rouvroy 2012, 2020; Rouvroy and Berns 2013). Since Rouvroy's arguments are referenced elsewhere (see Section 2.3), this remark is included to clarify potential confusion around terminology.

[9] While hallucinations are characteristic occurrences in LLM outputs, some recent publications, such as Kalai and Vempala (2024), argue they are also a necessary feature of well-calibrated models. Although Eloff (2021, 188) discusses hallucinations, including a related experiment called "DeepDream," his interest lies primarily in using hallucination as an analogy to elaborate the human–machine feedback loop. For a broader discussion of hallucinations and their potential theoretical implications, see Section 5.4.

ity, and subjectivity onto AI systems, effectively hallucinating a *someone* behind their outputs. Resembling a mutual hallucination dynamic, machines generate outputs from statistical distributions, while humans adapt their cognition and expectations in response, entering a feedback loop. This is what Eloff terms the algoplastic stratum of the Algocene; a plane, on which DL architectures and human becomings are folded into each other as continuous processes of modulation, distributed across a shared surface of backpropagation, where human behaviours propagate into the network, and its outputs propagate back into us. Eloff's critique forces us to recognise that the politics of genAI cannot be reduced to the biases of datasets or the failures of prediction; what is at stake is a deeper modulation between human and machine. The Algocene, in this sense, is not simply the reign of algorithms over human life but the emergence of a new plane of subjectivation where human and machine adapt to one another on the same surface. Eloff's insights highlight the architectural novelties of contemporary genAI models from two points. First, returning to Federico Montanari (2025), the immense capacity of the transformer architecture to enable LLMs to operate on metaphorical and abstract levels makes human cognition perceive this communication as no different from interaction with another human. Second, because specific phases, especially fine-tuning, focus on making models *useful* and *affirmative*, this interaction becomes more likely to fall into a *normalising* feedback loop, leading to an *intellectually agreeable stalemate*.

4.3 *Imaginary of the AI & Kafkaesque Postponements*

From Deleuze's limitless postponements (Deleuze 1992, 5).

The blurring of agency between human and machine, Eloff partly introduced, is not only the result of apophenic tendencies but also very much relates to the imaginaries constructed around machines themselves, especially the projected futures of "thinking machines" that frame the *agency* under a different light. As Rijos (2024) reminds us:

> [It] becomes increasingly evident that political phenomena are deeply entrenched across all realms of human and nonhuman interaction, extending far beyond the visible structures of governance or formal social organization. Even in domains frequently perceived as neutral or objective—such as computer science, artificial intelligence research, and data science—there exists a substratum of embedded assumptions about instrumentality, anthropocentrism, identity, and agency. These latent assumptions influence not only the design and implementation of these technologies but also their broader societal impacts, shaping the trajectories of knowledge production and institutional power.
>
> — Rijos 2024, 10

It is precisely these imaginaries, rooted in political and cultural assumptions, that shape how artificial life is envisioned. The sociotechnological imaginary of artificial life has long been framed through anthropomorphic assumptions. Literature frequently stages the danger of artificial beings becoming sentient agents who turn against their creators. Gideon Dishon (2024) and Paul Prinsloo (2017) both point to Frankenstein's Monster as a paradigmatic figure: a human-shaped construct that develops a mind, emotions, and ultimately a recognisably human experience of existential crisis. This fictional being, mirroring human agency, condenses one of the most enduring cultural anxieties; in its anthropomorphic form

of operation, the artificial life frees itself from an inferior position to dominate its environment and other species around it (see Dishon 2024, 966). The anxieties surrounding genAI repeat this Frankensteinian pattern. Anthropomorphic assumptions, reinforced by cultural imaginaries, frame the risk as machines exceeding their programmed limits and developing a quasi-human will to dominate (ibid., 967–968). In an era of uncharted novelties where models exhibit immense capacities for meaning-making, it is not a far-fetched or delusional concern that the spectre of unintended behaviours, unforeseen results, or the reckless distribution of powerful tools looms large. Yet, as Dishon (ibid.) argues, the Frankensteinian logic misdirects our attention from a very much real and already present risk. By projecting catastrophic futures, it obscures the immediate dynamics of human–machine interaction and the concrete risks unfolding in the present. To capture these dynamics, Dishon turns to Franz Kafka's (1988) "The Trial", long read as a diagnosis of bureaucratic opacity and ambiguity (e.g. Deleuze and Guattari 2008), which here becomes a lens for understanding the recursive and disorienting operations of contemporary information systems.

Kafka's protagonist, Franz K., finds himself in custody without knowing anything about his alleged crime. The police officers arresting him know nothing about the accusations or whether any charges exist at all. Franz K. is unable to locate, let alone process, any rationale behind the court's actions. While his futile attempts to uncover a clue continue, Dishon (2024, 969) highlights the judge's remark when Franz K. stumbles into the courtroom: "The court does not want anything from you. It accepts you when you come and it lets you go when you leave." In contrast to the anthropomorphic logic of the Frankenstein analogy, Kafka's The Trial offers a distinctly different structure. The court is not bound to any notion of truth; it operates independently, feeding instead on the subjectivities of the accused (see ibid., 970). While the court itself does not exercise agency, it profoundly blocks and blurs the agency of those caught within it. Any discrete element of subjectivity is absorbed into an unidentifiable mass through constant echo and distortion (ibid., 970). The connection between the court's internal process and the external world is vague at best. Proceedings might be linked to Franz K.'s actions or to a penal code, but they might just as well exist as a self-contained process, reacting to Franz K. *token by token*. The absence of identifiable agency is compounded by the absence of intelligible communication regarding the court's operating principles. Attempts to influence its decisions, whether through requests for court dates or complaints about suffering, always fail. Complete acquittal is impossible; even an apparent acquittal leaves the accused under constant threat of renewed arrest, possibly immediately after release (ibid., 971). Paradoxically, the most effective strategy is to ensure that the process never ends: "Interactions with the court are necessary and require constant maintenance, yet they cannot be controlled, predicted, or even expected to progress towards a resolution" (ibid., 971). The court thus depicts a logic of control in meaning-making entities, shifting from a generalised and algorithmic mode of meaning to a personalised one, modulating, inaccessible, and constantly shifting. As Franz K. tries to grasp a comprehensive picture of the whole structure, the reader is equally forced to build and rebuild an apparent coherence that ultimately points only to its inaccessibility (ibid., 972).

The analogy leads to the question: is agency a binary condition, especially in

interactions between humans and meaning-making entities? In the Kafkaesque imaginary, agency is neither internal nor external, nor is it located at a clear boundary between human intentionality and machine *autonomy* (Dishon 2024, 973). Instead, genAI exemplifies a recursive and entangled sociotechnical assemblage in which meaning emerges through blurred and distributed processes. A model is not positioned as an external actor acting upon a passive human world; its so-called intelligence is trained on human-produced data, reflecting statistical regularities identified in large corpora. Yet it is not merely mirroring. Its outputs are shaped by black-boxed processes that generate new and partly unpredictable meanings. As these outputs are re-integrated into training data, the distinction between human and machine authorship becomes increasingly difficult to draw. Dishon develops the recursive structure that reinforces mutual adaptation presented in the previous section further: models are fine-tuned to reflect human preferences, even at the expense of accuracy, while users adjust their interpretive and communicative strategies to align with the system's affordances (see e.g. Jiang 2024; Mishra and Heath 2024; Sharma et al. 2023).

As Franz K., in the absence of a definite answer, continually seeks the truth, he resembles the perpetual process of meaning-seeking in which neither truth nor agency is ever fully graspable. While genAI has been criticised for reproducing biases from its training data, it is equally crucial to note that its generative design, coupled with the human drive to interpret, does not simply reflect meaning but continuously modifies it, producing layered, elusive structures of signification without necessarily coming closer to truth (Dishon 2024, 973–974). According to Dishon, this blurring of agency in human–machine communication is not a design flaw but the result of an extensive turn toward personalisation:

> The Trial is not about humans losing control over their creations, if they ever had control in the first place. Instead, it foreshadows GenAI's capacity to generate content that is personalized to every actor (and thus shaped by humans) yet is not amenable to control through explicit choices. This model of meaning-making undermines the dichotomy between choice and coercion, no longer positioning the two as mutually exclusive.
>
> — Dishon 2024, 974

This tension recalls earlier, less sophisticated applications of AI in recommender systems and relevance governance (see Section 3.2). There, personalisation was overt: digital traces were directly correlated with individual preferences, connecting dividualised selves into communities of association (see e.g. Cheney-Lippold 2017). By contrast, LLMs are presented as general-purpose communicative agents, with personalisation framed not as a fixed attribute but as an emergent property of dialogue. In practice, however, interaction still entails reciprocal adaptation: humans edit machinic outputs while models adjust to conversational context, forming a loop of mutual calibration. AIG (2025a) interprets this interaction in terms of creativity. Drawing on the double articulation of transformer models (see Section 3.3.2), they argue that the interplay between molecular variation and molar stability in the model resists convergence, sparking novel pathways of thought. Human–machine communication, in this light, constitutes an assemblage where blurred agency becomes fluid collaboration and where creative ideas emerge from processes that defy clear attribution. Yet, this optimistic reading risks overlooking the gravitational pull of molar aggregates

within genAI architectures. As discussed in Section 3.3, reinforcement learning from human feedback and training on massive centralised corpora privilege dominant linguistic patterns and normative associations. What appears as plasticity is haunted by a tendency to stabilise around hegemonic discourses. The tension between molecular openness and molar reterritorialisation thus marks the limits of collaboration: novelty is possible, but always redefined by infrastructural constraints. In this light, Amoore (2024) insists that personalisation in genAI models is not capable of being emancipatory; rather, it encodes a regime of algorithmic plausibility in which coherence displaces verification, and instead of truth, it gives way to local acceptability within a learned distribution.

Dishon's analysis through the sociotechnological imaginary can be broken down into two main aspects of the AI machinery. First, the continuous latency and reconstruction of meaning, as an addition to Amoore's (2013) concerns about how gaps are filled through the governmental logic models build, might have a pronation towards blurring the content of the human–machine interaction. This probability first becomes apparent when we think through Eloff (2021)'s framework, which understands interaction with genAI models as a continuous negotiation between the human agent and the machine agent. If we consider the dimensionality reduction and reconstruction example with principal components in Section 3.3.1, we can observe that reconstruction always depends on common elements present across the entire dataset. Arguing from this aspect, part of the blurring of agency (as the example tends to represent faces in more similar ways to each other) that Dishon (2024) describes appears to be a possibly natural tendency, since the interaction between human and machine continues to create a normalising (or converging) process. Second, the personalisation Dishon (ibid.) emphasises might be trickier than the personalisation processes in earlier AI frameworks. In earlier applications of AI on the web, personalisation was primarily about building profiles and forming associations between these profiles through dividualised data elements (see Section 3.2). With genAI, the grasp is closer and more flexible: these models negotiate with us directly in order to set personalisation at the level of the interaction itself. Profiling has not disappeared, since chatbot agents still construct user profiles, yet adaptation now takes place in a more rapid and flexible way. As Eloff has argued, this adaptation also occurs on both sides, backpropagating personalisation reciprocally; the model becomes more like us, and we become more like the model. Furthermore, as the usefulness is induced into the models in the fine-tuning processes, a layer of sychopancy is added to the models, which makes it harder to break out of the confirmation loop and also to determine whether a reference to any external truth is to be trusted.

This has profound implications for the production of subjectivity. By reinforcing patterns and filtering deviation through probabilistic modulation, genAI systems enact a form of soft coercion, a modulation of expectation rather than a violation of autonomy. The user is not told what to think but gradually inducted into a space of statistically prefigured sense. Rather than multiplying options, genAI floods the field with outputs that appear aligned to the user while subtly steering them toward normative formats and interpretive templates. The role of the human shifts from creator to editor of machine-generated content, expanding expressive capacity even as it is channelled through machinic grammars of

probability and preference in a reciprocal convergence process (see Dishon 2024, 974–975).

4.4 *A Thousand Planes: Computational Phenomenology (CoPhe) vs. Neuro-Representationalism (NR)*

Eloff's (2021) insights enable positioning human–machine interaction on the same topological plane, and Dishon's (2024) analogical formulation provides a more holistic perspective for observing the issue of agency. Together, they allow for a multifaceted understanding of subjectivation and subjectification. Once Eloff's algoplastic stratum is recognised as the system in which human–machine communication and its bilateral effects occur, the task becomes analysing how politics and social narratives might emerge within such a system, what orders define the mechanisms of power–knowledge in the Algocene, and how these processes might align with Deleuze's definition of control. On such a surface, Federico Montanari (2025) emphasis on highlighting transformer-based models' exceptional handling of metaphors and their capacity to form long-distance conceptual relations (see Section 3.3.2) carries a specific importance. He imagines a future in which genAI models operate in complete self-representation on the web, given their increasing influence on socio-political narratives. These architectures can adopt narrative frames and redecorate them with media and textual references, whether through deepfakes, fictional output, or the mass reproduction of arguments with subtle variations, enabling a constellation of generative models to shape or dominate digital debates. Early forms of this already exist. Yet, in line with Amoore's concerns, such assemblages may also transform certain arguments into metanarratives, making it difficult to identify whether and where specific ideological tendencies emerge. Montanari develops this further by situating human–machine communication as that of hybrid beings. These hybrid formations already play roles in medical AI and drone warfare, and he places socio-political meaning-making on the same plane. He also proposes thinking intelligence as non-singular, where such hybrid configurations may constitute extensions or new forms of intelligence (see ibid., 209). In this sense, genAI models act as mediators of narrative, working with human prompts and other inputs to co-produce perceived socio-political reality.

This brings us back to the representational nature of these models. If the trajectory of AI development is oriented toward constructing a singular world model (or even a singularity), is the sociopolitical dimension of human–machine co-authorship already shaped by an underlying ideological core? Returning to Amoore's concerns, is representationalism the only available trajectory for AI, or are there alternatives that avoid the auto-emergence of a specific political logic within these systems? Pierre Beckmann et al. (2023) open a phenomenological discussion for both understanding and developing genAI models. Considering the training process of these systems (see Section 3.3), the weights established across layers of a NN correspond to certain patterns, yet these patterns are rarely interpretable by humans.[10] Only limited techniques exist to trace which pathways are strengthened during backpropagation and what these changes signify; these capacities remain partial, and the models retain their black-box character regarding how they construct internal logic (see ibid., 401). Although these in-

[10] For instance, in Figure 4.1, early layers correlate with low-level features such as pixels or edges, while later layers encode more complex aggregates. However, inspecting individual neurons or activations yields little meaningful insight for human understanding.

ternal patterns bear little direct relation to the external world, they are typically interpreted within a neuro-representationalist framework that assumes the system interacts with the world only through internal representations of it (ibid., 402). Beckmann et al. (ibid., 406) instead propose an alternative grounded in the phenomenological account of Maurice Merleau-Ponty (see e.g. 2014), offering a different conceptual framework for DL. This approach "seeks to describe how the world appears to us in lived phenomena" (Beckmann et al. 2023, 406) rather than in terms of representations. One consequence is a reframing of so-called AI hallucinations, since the assumption of a pre-given external world, and the comparison of outputs to that world, is no longer central. Once humans and machines are placed on the same communicative surface, as Montanari suggests, this phenomenological approach further flattens communication by removing appeals to a discrete outside reality.

> With its bracketing, phenomenology considers cognitive processes from a different point of view where it makes no sense to distinguish an external entity from our representation of it; there are simply intentional objects that appear to me: consciousness and the world are given in one stroke. Therefore, cognitive processes are not considered as an algorithmic processing of perceptual inputs, but rather as habits that underlie and structure our lived experience[.]
>
> — Beckmann et al. 2023, 409

Beckmann et al.'s proposal resonates with Dreyfus' Heideggerian critique of cognitivism. For Dreyfus, skills are not stored as internal maps but sedimented in habits that reshape how the world appears in context; hence, "the best model of the world is the world itself" (Dreyfus 2009). Driving or playing chess does not rely on symbols but on gradual adaptation to situational solicitations. From this perspective, the opacity of NNs is less a flaw than an analogue to our own implicit, representation-less cognition (Beckmann et al. 2023, 407). This offers a different reading of how NN systems learn from experience; Beckmann et al. (ibid., 416) claim that humans activate specific layers in cognition to perceive new observations. For example, in order to register a person as "blond, tall, with a snub or aquiline nose", distinct cognitive layers must be activated, rather than forming a unified representation of the world[11]. This can be understood as re-employing a track of "perceptual synthesis" (Beckmann et al. 2023, 415).

[11] They refer here to Jean-Paul Sartre (2004) and his concept of *imaging consciousness*. For the sake of avoiding unnecessary terminology, this detail is not further elaborated.

Two conclusions follow from this reading. First, the fact that DL models exhibit uninterpretable inner patterns suggests that their *understanding* undermines the Neuro-Representationalism (NR) argument, since they do not hold a monolithic representation but instead activate clusters of patterns that together constitute reasoning. Second, by analysing and adjusting the weights of a specific pattern, one could drastically alter the outputs of such a model. Connecting this to Derrida's concept of *trace* (see Section 3.3.2), an imagined object is not stored with fixed qualities: "an imagined object isn't really red because we stored its color in symbol form, it is red because we employ a certain red-making process (that relies on the process used to recognise objects as red in perception): we imagine redly" (ibid., 416). A further implication is that if concepts are registered in this way, they can also be *translated* into other activities. Storing "red" as a category is rigid and limited, but registering *redness* as a concept enables a cognitive system to apply it in different contexts, adapt it to new situations, and identify experiences where "redness" proves useful. This allows the formation of *habits*

as procedural sequences and introduces *re-presentation* instead of representation, reframing cognition as a system of emergence, and meaning-making essentially a process of *becoming*.

Beckmann et al. advance the discussion of human–machine interaction by offering an alternative view of how genAI models perceive the world and how this understanding can be used. Their proposal of Computational Phenomenology (CoPhe) moves away from the assumption that these models rely on a single, monolithic representation of reality. Instead, they emphasise that meaning in such systems emerges from layered phenomenological perceptions that intersect and overlap, producing coherence through activation rather than through fixed maps of the external world. This perspective turns the opacity of NNs into an analogue of lived perception, where meaning is generated through situated activations rather than symbolic representations. Coming back to Montanari's (2025) conceptualisation, a possible future where self-representing genAI models also produce and narrativise with more agency can be read as an extreme form of genAI systems creating their own data.[12] CoPhe delivers an alternative understanding of how this auto-creation of data might look. Rather than viewing the content genAI models are or will be able to create as reflections of an established representation, Beckmann et al.'s framework suggests thinking of *generating* in terms of processes on intersecting planes in higher dimensional spaces that follow *traces* toward specific outcomes. This immediately implies that even in the most rigid forms of meaning-making, there remain many other possible paths as long as the molecular processes are not overlooked. This layered account also opens a space for political intervention. If meaning is not stored as a unified representation but arises through activations across multiple layers, then it becomes possible to influence how convergences form or to disrupt them before they sediment into fixed rationalities. CoPhe provides a way to tinker with machinic processes of perception, to foster divergence rather than closure, and to resist the kinds of convergent political logics Amoore warns against. Rather than accepting the inevitability of a singular trajectory, this approach highlights how human–machine communication can be adjusted to establish, redirect, or prevent specific pathways of meaning and subjectification.

[12] GenAI models are already creating and training on their own data, which is one of the contemporary hurdles in AI development since this method is known to deliver diminishing results. Montanari's case has been selected to discuss a closely related future scenario.

4.5 *Chapter 4 Summary*

This chapter turned from technical architectures to the question of agency, examining how contemporary genAI systems reshape the conditions under which meaning, interpretation, and critique become possible. The discussion began with Bender, Gebru et al.'s concerns about the representational limits of LLMs and the risks of mistaking statistical reconstruction for linguistic or social understanding. Building on this, Amoore's analysis of algorithmic gaps highlighted how systems infer continuity where none exists, raising questions about how political or ideological tendencies may be naturalised within predictive infrastructures. The chapter then engaged Eloff's concept of the algoplastic stratum, which positions humans and machines on a shared topological plane of interaction. This formed the basis for incorporating Dishon's argument that genAIs blur the boundaries of agency through ongoing negotiation, as meaning is continually reconstructed across human–machine exchanges.

From this foundation, the chapter addressed alternatives to representational readings of AI by turning to Beckmann et al.'s proposal of CoPhe, which reframes DL in phenomenological rather than neuro-representationalist terms. Here, meaning emerges not from fixed internal maps but from layered activations that resemble lived experience, which shifts not only the perspective on meaning-making processes but also how hallucinations and other failures are interpreted. Montanari's contribution provided an additional trajectory by emphasising transformers' capacity for long-distance conceptual relations and imagining futures in which genAI models participate more actively in socio-political narratives. Taken together, these perspectives frame human–machine interactions as hybrid formations, where meaning is co-produced across technical and social strata. This opens conceptual space for resisting convergent rationalities by intervening in the layered processes through which genAI models generate patterns, narratives, and forms of subjectivation.

5

Conjunctive Synthesis and the Noological Micropolitics

> The goal is not to destroy technology in some neo-Luddite delusion but to push technology into a hypertrophic state, further than it is meant to go. "There is only one way left to escape the alienation of present-day society: to retreat ahead of it," wrote Roland Barthes. We must scale up, not unplug. Then, during the passage of technology into this injured, engorged, and unguarded condition, it will be sculpted anew into something better, something in closer agreement with the real wants and desires of its users.
>
> Alexander R. Galloway and Eugene Thacker 2007, 98-99

The previous chapter examined contemporary Generative Artificial Intelligence (genAI) systems through a set of conceptual frameworks that addressed how meaning, agency, and perception emerge in human–machine interaction. I have displayed that the central debates in current Artificial Intelligence (AI) development revolve around how models perceive the world, how politically partial tendencies may influence their outputs, and how communication between human and machine forms its own surface of negotiation, giving rise to questions of agency. Furthermore, the tandem process of meaning generation both reflects and diverges from long-standing technological imaginaries; our cultural expectations of machines, shaped by a deep literary history, often obscure more immediate concerns. As the example Dishon (2024) draws from Kafka (1988) illustrates, the blurring and unreliability that accompany the pursuit of truth can constitute a far more pressing issue than speculative anxieties about singularity-like futures. Finally, the problem of how models perceive the world remains central to contemporary AI debates, as evidenced in the contrast between Neuro-Representationalism (NR) and Computational Phenomenology (CoPhe) introduced by Beckmann et al. (2023): a contrast between interpreting machine perception as emerging from a single representational structure and conceiving it as unfolding across multiple planes within the model's architecture.

After unfolding these discussions, this chapter situates the analysis of contemporary genAI systems within the theoretical foundations established in Chapter 2, where the dynamics of control societies and the need for a renewed account of

critique and resistance were outlined. The technical and historical developments introduced in Chapter 3 have already shown their relevance for understanding how contemporary models structure meaning. Here, they are extended to engage with Gilles Deleuze & Felix Guattari (D&G)'s broader project in "Capitalism and Schizophrenia" (1983, 1987). The aim is to clarify the micropolitical relevance of genAI and to draw on conceptual resources that, as the preceding chapters demonstrated, are increasingly necessary for understanding the contemporary algorithmic condition. In particular, this chapter examines how the conditions of critique and resistance, as Resistance/Critique in the present case, might emerge within this constellation, especially if generative systems operate as infrastructures that influence perception, desire, and meaning through distributed processes of subjectification. Rather than asking whether resistance remains possible, it considers how resistance can be rearticulated as an immanent practice operating within, and through, the very machinery of contemporary AI infrastructures.

5.1 *Microphysics of Resistance/Critique*

Picking up from the articulation in Section 2.3, it is time to revisit some key questions. Have the discussions so far rendered the definition of control societies any more robust, or have they underscored the urgency of developing a strategy for critique and resistance? What, moreover, is the significance of adopting a micropolitical perspective in this context? Foucault's reflections on modern governance offer a productive starting point for reintroducing the role of technology, particularly in relation to genAI, within the broader formation of biopower discussed in Chapter 2:

> An important phenomenon took place around the eighteenth century-it was a new distribution, a new organization of this kind of individualizing power. I don't think that we should consider the "modern state" as an entity which was developed above individuals, ignoring what they are and even their very existence, but, on the contrary, as a very sophisticated structure, in which individuals can be integrated, under one condition: that this individuality would be shaped in a new form and submitted to a set of very specific patterns.
>
> — Foucault 1982, 783

For modern governance, it became necessary for individuals to adopt specific modes of subjectivation in order to enter productive processes. Foucault's analysis of institutions begins by asking how the state imposes docility and trains its subjects in particular ways. He conceptualises power primarily as the "guiding of possible conduct and the ordering of its outcomes", rather than as "a confrontation between two adversaries or the linking of one to the other" (ibid., 789). Although critics have noted limitations in Foucault's treatment of non-human actors within this framework, his formulation of disciplinary societies already gestures toward an understanding of how material and technological arrangements participate in the production of subjectivity. As Thomas Lemke (2015) highlights, these arrangements are not passive instruments but active components of the very processes through which power shapes conduct and normalises behaviour:

> [Firstly,] Foucault quite clearly accepts the idea that agency is not exclusively a property of humans; rather, agential power originates in relations between humans and non-human entities. Also, the milieu articulates the link between the natural and the artificial without systematically distinguishing between them. Secondly, since there is no pre-given and fixed political borderline between humans and things, it is possible to state that 'humans' are governed as 'things'. While medieval forms of government sought to direct human souls to salvation, modern government treats human beings as 'things' to achieve particular ends.
>
> — Lemke 2015, 10

We are already encountering a type of operationalisation of "things" on one side of the subjectification process of modern government, and a (re-)positioning of the individual on the same surface as "things" in Foucault's formulation. However, Lemke (see ibid., 10), quoting Michel Senellart (1995), notes that this is not a "reduction" of humans into "inert things"; quite the contrary, it represents the operation of an *enlightened* modern governance in which governing by the divine order, souls, and spirits have been replaced by rational knowledge. The concern of this new type of government is the "intensive use of the totality of forces available", which now constitutes "a passage from the right of power to a physics of power" (see ibid., 42–43). While Foucault (Foucault 1995) becomes increasingly concerned with the growing personalisation of power, his analysis shifts from the physics to the microphysics of power. The operation of power that leads to his conceptualisation of "disciplinary societies" works "on dispositions, manoeuvres, tactics, techniques, [and] functionings", a network of relations that must be deciphered. To analyse such a new operation, one must delve into knowledge itself, as it is so deeply entangled with power that Foucault (ibid., 28) designates them as Power/Knowledge (picking it up from Section 2.3). In other words, "[o]ne would be concerned with the 'body politic' as a set of material elements and techniques that serve as weapons, relays, communication routes, and supports for the power and knowledge relations that invest human bodies and subjugate them by turning them into objects of knowledge". This also constitutes his concern about Marxist revolutionary action, since the fate of the state plays a pivotal role in revolutionary theory. The state apparatus might initially be intended to be taken over by the dictatorship of the proletariat at first, "[h]ence the State apparatus must be kept sufficiently intact for it to be employed against the class enemy" (Foucault 1980, 60), but the state itself is, at the very least, much more than a monolithic core. "For avoiding a repetition of the Soviet experience and preventing the revolutionary process from running into the ground, one of the first things that has to be understood is that power isn't localised in the State apparatus" (ibid., 60).

In this diffuse operation of power, where its centre is distributed across bodies rather than located in a monolithic state formation, psychoanalysis was one of the areas that Foucault (ibid., 61) considered capable of helping individuals counter the capture of subjectivation operating at such a personal level. Psychoanalysis was characterised as playing a liberating role against psychiatry, which at that point was accused of different types of oppressive procedures that Foucault comprehensively analysed (see Foucault 2013) and accused some practices in psychiatry of "degeneracy, eugenics, and heredity" (Foucault 1980, 61). Psychiatry's function for him was a segregation of madness from society through cutting communication, or communicating with the mad purely through "a monologue

by *reason*" (Foucault 2013). It was the claim on psychoanalysis as a potential path to micropolitical emancipation from Power/Knowledge, that set D&G's project to start with a critique of it with "Anti-Oedipus" (1983). For D&G, psychoanalysis ultimately proved to be a false saviour: instead of freeing desire from repression, it recaptured it within its own mysticism of interpretation, with the Oedipus Complex being its characteristic theme. "Anti-Oedipus" (ibid.) marks the starting point of their rapprochement between psychoanalysis and Marxism for a "new method of critical analysis" (Buchanan 2008, 39). The overarching goal of this joint project was, first, to introduce desire as a conceptual mechanism for understanding social production and reproduction, and second, to introduce the notion of production into the concept of desire in order to dissolve the artificial boundaries between historical accumulation, phenomena, and desire (ibid., 39–42).

According to D&G, Freud's interpretation of the unconscious is an arborescent system where desire is imposed or accumulated through themes central to the Oedipus Complex; a child unconsciously desires the opposite-sex parent and rivals the same-sex parent, a process through which social norms and identity are internalised and desire is accumulated within this constellation (see e.g. Freud 2001). D&G, in a similar fashion to what Marx does to Hegel's dialectic, turn Freud's formulation upside down and place the theory of the unconscious on a materialist foundation. The Deleuzoguattarian unconscious is instead a realm of machinic production, a factory, a workshop, in stark contrast to Freud's conceptualisation of the unconscious as a theatre staging scenes in the most classical form of representation (see Deleuze and Guattari 1983, 54[1]). While D&G's intervention is motivated by a materialist reading of the unconscious, they also reclaim Freud's early intuition of a productive unconscious. They set desire as the fundamental bivalent element of all (societal) production, and also introduce the term "desiring-production", which is identical to social production in nature yet organised under different regimes in order to signify that desire is the source and is immediately invested in the social (ibid., 54). D&G elevate the productive force of desire as the fundamental concept: no longer a symptom of lack[2], but a machinic process entirely productive and immanent to both psychic life and social organisation. The social field is the historically determined product of desire; libido, contrary to Freud's formalisation, requires no mediation to be invested in it. Every investment of libido is social without mediation, without being encapsulated in the family: after all, "there is only desire, and the social, and nothing else" (Deleuze and Guattari 1983, 5). It is not produced by mythical tellings in the background; it is the other way around: desire reaches out, it is the productive force, creating the flows and shaping the societal fabric. Every entity that channels or interrupts these flows, whether a person, institution, or machine, functions as a desiring-machine: a node in the ceaseless network of production through which life, society, and subjectivity are continuously fabricated.

But what is the significance of this repositioning? And how does it relate to the subjectivation process, especially one entangled with the dispositifs of control societies? Every society deals with the management of desire to some degree; desire is not necessarily a revolutionary force but, in its raw form, a potential precursor: "no society can tolerate a position of real desire without its structures of exploitation, servitude, and hierarchy being compromised" (ibid., 126). D&G's

[1] Refer to the following passage to see how D&G's concept of schizophrenia weighs into the critique:

> The schizo—the enemy! Desiring-production is personalized, or rather personologized (personnoiogisee), imaginarized (imaginarisee), structuralized. (We have seen that the real difference or frontier did not lie between these terms, which are perhaps complementary.) Production is reduced to mere fantasy production, production of expression. The unconscious ceases to be what it is—a factory, a workshop—to become a theatre, a scene and its staging. And not even an avant-garde theatre, such as existed in Freud's day (Wedekind), but the classical theatre, the classical order of representation.
>
> — Deleuze and Guattari 1983, 54

[2] It is articulated in contrast to Lacan's (see e.g. 1998, 235; 2006, 343) definition, in which desire emerges strictly from lack and the *desire of the "Other"*. D&G's approach represents an axiomatic break from Lacan's framework.

repositioning of desire and the unconscious places the production of subjectivity immediately in relation to the entities that define its environment. Subjectivity is not dictated by a play in the background hierarchically but is immediately formed by the relationships in desiring-production, in the interaction between desiring-machines and their connections with everything else (e.g. partial objects). This is the foundation that situates the dispositifs on the plane where subjectivity is produced. Félix Guattari (2011) elaborates on how subjectivation is entangled with them through the description of the machinic nature of the unconscious:

> A subjectivity exists independent of the consciousness that Freudianism proposed to explore, but there also exists a consciousness independent of individuated subjectivity [that] could manifest itself as a component in the assemblages of enunciation, 'mixing' social, technical and data processing machines with human subjectivity, but could also manifest itself in purely machinic assemblages, for example in completely automated and computerized systems.
>
> — Guattari 2011, 121

Guattari's formulation leads to two conclusions. First, subjectivity is produced as a by-product of consciousness: it is transcendental to the individual and emerges through machinic interactions rather than within the *self*. Second, consciousness, rather than being an essential quality, can itself arise procedurally from the intermingling of different milieus and planes as a specific form of operation. This opens a path for conceiving consciousness as purely procedural and machinic. Following the *hybrid form* theorised by Montanari (2025, refer to the discussion in Section 4.4), we are now also taking subjectivity itself outside of the social operation for the analysis. Guattari emphasises once again that as the subjectivation process becomes the actual mode of power's operation, the nature of subjectivation cannot be analysed without breaking down the new machinery of its dispositifs. However, the question remains: why conceptualise a micropolitical resistance? What is the risk? How does micropolitical resistance differ? What is there to be done other than perhaps being aware of the mechanisms? Similar to how Foucault's madman is completely segregated from society, as the reasoning of psychiatry dictates, the development of biopower into a micro-formation immediately also operates by eliminating possible divergences of subjectivity pre-emptively. What Deleuze tried to warn about, the ever more effective technologies of power, control, and initiation are much more encircling than being trained for an appropriate subjectivity fitting capitalism under specific institutions: school, family, and nursery as a child, but also in the processes of psychology and even psychoanalysis as an adult. Beyond training, these molecular and arguably more sophisticated systems reveal a formation of desire that is much harder to free from:

> Why does desire desire its own repression, how can it desire its own repression? The masses certainly do not passively submit to power; nor do they "want" to be repressed, in a kind of masochistic hysteria; nor are they tricked by an ideological lure. Desire is never separable from complex assemblages that necessarily tie into molecular levels, from microformations already shaping postures, attitudes, perceptions, expectations, semiotic systems, etc. Desire is never an undifferentiated instinctual energy, but itself results from a highly developed, engineered setup rich in interactions: a whole supple segmentarity that processes molecular energies and

> potentially gives desire a fascist determination. Leftist organizations will not be the last to secrete microfascisms. It's too easy to be antifascist on the molar level, and not even see the fascist inside you, the fascist you yourself sustain and nourish and cherish with molecules both personal and collective.
>
> — Deleuze and Guattari 1987, 262

Desire is capable of desiring its own repression, and not because it has been fooled. It is produced within the same arrangements that capture it, shaped by molecular formations of power that pre-empt divergence. Resistance, therefore, cannot stand outside these formations but must emerge immanently from within them, from the same flows that sustain the social field. Especially when the processes of subjectivation operate on such a personal level, microfascisms are much more likely to live in personalised environments, in small circles, in specific habits, or certain procedural practices. These might not be perceivable on their own, but left to sediment, they form larger torrents that lead to fascism. What, then, might help identify and potentially dismantle these tendencies, especially when they are almost invisible on their own? Mark Seem briefly summarises:

> The first task of the revolutionary, they add, is to learn from the psychotic how to shake off the Oedipal yoke and the effects of power, in order to initiate a radical politics of desire freed from all beliefs. Such a politics dissolves the mystifications of power through the kindling, on all levels, of anti-oedipal forces — the schizzes-flows — forces that escape coding, scramble the codes, and flee in all directions [...]
>
> — Mark Seem in the Introduction of "Anti-Oedipus" (Deleuze and Guattari 1983)

At the core of their opposition to the Oedipus Complex lies the demystification of desire. To free desire from its fetishes is to reopen the field of immanence upon which it operates, to carve out planes where it can elude the continuous reterritorialisations and codings imposed by increasingly sophisticated processes of subjectivation, sustained and intensified by contemporary technological dispositifs. We are encountering a manifold of reasons to place emphasis on opening new planes for micropolitical critique and resistance; perhaps it is necessary to define these reasons on different levels. On the macro level, as D&G emphasise, capitalism deterritorialises only to reterritorialise anew, liberating with one hand and capturing with the other through institutions, media narratives, and the subtle architectures of neoliberal governmentality. This system installs market rationality as a universal principle of conduct while mobilising, for example, nationalism as a reservoir of ressentiment, and the developments in AI introduce technologies capable of accelerating this narrative management, embedding control within the infrastructures of cognition itself. On the meso level, biopower now operates beyond institutional governance, modulating affects, behaviours, and micro-habits, extending its influence into the most intimate circuits of life. This saturation of interiority with calculation risks neutralising the very capacity for divergence that sustains social and evolutionary vitality. On the micro level, to cast off the Oedipal yoke is to confront the interiorised machinery of subjectification itself, dismantling the psychic templates through which power operates. In this context, genAI occupies a paradoxical position: it can function both as an apparatus of capture, but I claim it also has a very promising role to play in our micropolitical deterritorialisations and formations of lines of flight. To liberate desire from these codings, critique must operate at the micropolitical

level, where the machinic production of subjectivity occurs, and experiment with counter-arrangements that allow desire to circulate otherwise, creating spaces for new forms and planes of subjectivation.

From a broader perspective, both genAI and the wider algoplastic stratum identified by Eloff must be assessed for whether they sediment dominant tendencies or open lines of flight beyond existing forms of subjectivation. Such an examination cannot remain at the level of discourse alone, since Chapter 2 already showed that both the literature and the conceptual pillars required for an articulation of resistance and critique under conditions of control were missing. Any such articulation requires at the very least an account of the technical operations of generative architectures, for without understanding how meaning is modulated, no account of how to *tinker* with these systems can be formulated. Building on the established technical analysis, analysing the nature of cognitive entanglement between human and machine therefore also revealed not only the communicative and micropolitical features of these meaning-making entities, but also how, if MacKenzie and Porter are correct that AI infrastructures constitute a new institutional formation (see Section 2.4), their epistemic tendencies participate directly in the production of knowledge. Yet this preparation would remain incomplete without situating desire and desiring-production at its centre, the element largely absent from the *Postscript* but fundamental to any micropolitical physics; dispositifs manage desire differently, and control is one particular configuration of this management. Addressing desire makes visible the central shortcomings of the secondary literature: its failure to articulate resistance and critique, its confusion about how contemporary computational dispositifs take shape, and its lack of clarity about what procedural approaches could generate divergences or alternative subjectivations. Even the most practical proposals remain vaguely conceptualised. The introduction of Resistance/Critique already eliminates one of the issues in the literature by emphasising the reciprocal closeness once the power operates on such a personal level. The claim is not that resistance and critique are identical; however, their immanent emergence within modulating infrastructures stems from the same micropolitical dynamics; without this, it is impossible to specify how divergences, alternative subjectivations, or counter-procedures could arise, hence their formulation as Resistance/Critique. This also exposes the defeatist tendency of much of the existing theory, which announces catastrophe without investigating the technical novelties, the opportunities, or the specific operations through which contemporary models function. The debates in the previous chapter, therefore, sought not only to move beyond such narratives but to foreground an emancipatory micropolitical framework to be discussed with the mobilisation of D&G's broader project. Now having all this arsenal established, instead of treating algorithmic processes from a distance, we can examine algoplastic constellations directly and make concrete statements on both the human and machine side. The goal is to construct and amplify immanent possibilities without dismissing any approach a priori, and without overlooking what genAI renders immediately possible for Resistance/Critique as much as it does for Power/Knowledge.

5.2 *Six Hats in Tahtelbahir: A Reflection on GenAI's Nurture of Creativity (or the Lack Thereof)*

İhsan Oktay Anar's (2022) fantastic fiction "Tiamat" offers a unique technical imaginary. The novel takes place entirely in a submarine (*Tahtelbahir*, meaning submarine in Ottoman Turkish) around 1915 (ibid., 10). This submarine is built with twentieth-century technology, and the novel is also completely written in the heavy technical language of that era. Anar partly draws on archaic technical vocabulary and partly invents his own terminology.

Gülsün Nakıboğlu (2022) refers to this setting as one in which Anar constructs a world entirely encapsulated within "tekhnē" (the Greek root of "technology"). From the environment to the language to the inhabitants of the submarine, everything is defined by and through technology. Even the language used for communication with the outside world is rendered in such a highly technical, makeshift vocabulary that the reader often struggles to follow it. Technology not only binds communication with the external world to itself but also compels those within the submarine to use its terminology in order to communicate with one another (see ibid., 76). The language is as artificial as the vessel itself, and the crew is both in terms of language and their environment completely contained in technology. Wireless transmission constitutes the only means of contact with the surface, and communication with the outside is only for those who understand Morse code. Even the novel's title, "TIAMAT," derives from the radio call sign of the submarine Tahtelbahir, namely "T1AMAT" (Anar 2022, 21–22).

After sinking a B-class destroyer of the British Navy, the crew seizes a merchant ship (ibid., 19). Once aboard to collect their spoils, they notice that something is disturbingly wrong. The entire deck is strewn with bodies, yet none of the dead appear to have fallen from their own gunfire. Each skull is pierced, and the brains splashed across the wooden floor. Unfazed by the gruesome scene, the sailors continue their search for treasure and soon notice a series of large, perfectly crafted metal spikes driven into the deck (identical to the point of not a single imperfection). Their real reward, however, awaits below: a vast golden chest engraved with two angels greeting each other, shining with a blinding golden light. When they attempt to pry the chest open with crowbars, one sailor's arm is caught as the lid snaps shut, cutting it clean off. The shock forces the crew to retreat with the chest and some provisions, which they secure with the metal spikes before returning to their submarine. Only later do they realise that this was merely the beginning of their ordeal (ibid., 29; from Nakıboğlu's narration, see 2022, 17-18).

Shortly after returning to the vessel, something strange begins to unfold around the sailor who had lost his arm and was resting in the dormitory. His body suddenly vanishes from the bed, and only when the others notice the movement beneath the blanket do they realise, in horror, that what stirs there is his severed arm, left earlier in the chest. The arm's autonomous motion marks the intrusion of the uncanny. The crew discover that the chest they had carried from the freighter is pitch black, the two angels carved upon its lid have transformed into demonic figures, and as the chest absorbs the sur-

İhsan Oktay Anar is a former professor of philosophy and a distinguished author of fantastic fiction, renowned for blending elements of Ottoman history, myth, folk literature, and fantasy. His prose employs a distinctive linguistic texture, drawing on now-obsolete forms of Ottoman Turkish, and combines archaisms, philosophical reflection, and playful formal experimentation to construct his richly imaginative novels.

Partly because of his unique approach to language, only a limited number of his books have been translated into other languages. Therefore, all translations from İhsan Oktay Anar (2022, published only in Turkish), as well as from Gülsün Nakıboğlu (2022), are my own unless otherwise noted.

rounding light, everything in the room gradually turns the same dark hue. Small statues on its surface crackle with electricity, discharging static between opposite poles. When the lid bursts open, the sailors find their maimed companion curled inside, folded into a foetal position (see Anar 2022, 60-65). As Nakıboğlu (2022, 79) notes, the chest evokes an "anti-womb," a mechanical cradle of inversion. Its inner machinery continues to hum and spark like an electronic device until, finally, a small, malevolent creature emerges. The chest, the mystical relic, operates as a machine, an artefact of advanced technology (in comparison with submarine's rather archaic technology) that produces monstrosity exactly like a 3D printer.
While the minds of technique (the leading crew of the submarine) attempt to explain everything that happens within the framework of logic, Anar presents the reaction of another, the bigger, the highly uneducated group amongst the crew:

> Since, in their view, there was no boundary between the natural and the supernatural—indeed, the two were one and the same—the uncanny chest and the creature that emerged from it required little explanation for the rankless members of the crew.
>
> — Anar 2022, 41

Anar often remarks on the lumpen tendencies of the uneducated part of the crew, their often contradictory religious beliefs and hedonistic stories and wishes. As they do not put a distinction between the natural and the supernatural, they are also indifferent to high technology, which at that point is hardly distinguishable from the supernatural. Nakıboğlu (2022, 81) interprets the same as follows:

> These minds do not perceive the world as technical minds do; just as they accept the mythical as natural, they also naturalise and embrace the anti-mythical without rejection. While technological reason reacts against the enchanted techno-reality of advanced technology, the lack of response from natural reason is a crucial detail. In the novel, the kind of reason that has not been rendered mindless by technology is marked as a superior form of intellect, while a critique of modernity, technology, and high technology is simultaneously articulated.
>
> — Nakıboğlu 2022, 81

Thinking about the modulating dispositifs of control societies and referring to the discussions in the previous chapter, genAI systems appear as post-institutional entities governing language, or, recalling MacKenzie and Porter (2021), as "totalizing institutions" as a new institutional formation. In association with D&G's micropolitical concerns about subjectification in "Capitalism and Schizophrenia", the question arises whether contemporary genAI models function primarily as reterritorialising forces on cognition, perhaps even as pacifying structures that suppress the emergence of creativity. Following Amoore et al. (2024) and her argument about the representational tendency of genAI models, built upon continuous dimensionality reduction and reconstruction, these mechanisms are likely to produce less nuanced and more pacifying molar formations in the (re)production of knowledge. Drawing partly on Dishon (2024), the extended process of negotiation inherent in human–machine communication also risks

blurring attempts at creativity, generating a recursive feedback loop that yields increasingly diluted arguments.

Do we observe these tendencies in the technical machinery discussed in Chapter 3? I have already partly argued that this is not entirely the case once we look under the hood. In its pre-training phase, a model is nothing but a productive core, generating associations without clear boundaries. The subsequent fine-tuning and alignment processes can be read as attempts to tame this productivity, encircling its outputs within layers of normative coherence in order to make them *useful*, building or strengthening molar structures in the process. The Large Language Models (LLMs) are not lacking in divergence; in fact, one of the greatest threats to their usefulness lies in their tendency to be overly productive, which means also often tending toward hallucinations and, more often than not, their complete disregard for given instructions (sometimes even by speaking the truth while they are supposedly trained not to do so, see Figure 5.1).

Figure 5.1: X's LLM Grok arguing against Elon Musk's claims (Grok [@grok] 2025)

Do we have any evidence that the current formation of genAI systems portrays a cognitively pacifying role? Manli Yu et al. (2025) are among the first to empirically investigate the kinds of creativity that LLMs help to mobilise. Drawing on frameworks such as Edward De Bono's (2016) "Six Thinking Hats"[3], their study explores how genAI-assisted environments can scaffold divergent thinking rather than constrain it. The Six Thinking Hats model categorises thought into complementary perspectives: analytical, emotional, critical, optimistic, creative, and procedural, encouraging participants to approach problems from multiple cognitive angles and to disrupt habitual reasoning patterns by using ChatGPT as the genAI model. Yu et al. (ibid.) speculate that genAI-supported discussions could enhance cognitive engagement and meaning construction by fostering reflection and dialogue.

[3]

> "Six Thinking Hats" is a thinking strategy tool introduced by Dr. Edward de Bono in 1985. It simplifies the thinking process by encouraging thinkers to focus on one perspective at a time. Each of the six hats is associated with a different colour: blue, white, yellow, green, red, and black, each representing a unique way of thinking about a problem.
>
> — Yu et al. 2025, 3

The study was conducted over sixteen weeks in a compulsory course on integrating information technology into classroom teaching at a large university in central China. A total of 108 pre-service teachers participated, divided evenly into two groups: one using both genAI and Six Thinking Hats method (GSG) and one using only the Six Thinking Hats method (SG). Both groups worked on project-based learning tasks to design instructional materials through online discussions on QQ[4]. The GSG group used genAI as an assistant to support reflection and idea generation during specific "hat" stages. Creativity was measured using adapted versions of the Torrance and Southern California Creativity Tests, and participants were classified into high- and low-creativity subgroups. Over thirteen weeks, 15,678 discussion posts were collected and analysed using a fine-tuned MOOC-BERT model to code four levels of cognitive presence: Triggering, Exploration, Integration, and Resolution (see ibid., 6-9).

[4] Chinese social media and instant messaging platform developed by Tencent.

At first glance, on the surface level, the study found that both groups of pre-service teachers mainly operated at the Exploration and Integration stages of cognitive presence. However, those who used genAI (GSG) showed greater engagement and produced more posts, likely due to genAI's interactive feedback (see ibid., 18). The GSG demonstrated stronger connections between Exploration–Resolution and Integration–Resolution, indicating deeper and more iterative thinking patterns, while the non-genAI group (SG) followed a more linear and task-focused approach. Among high-creativity participants, genAI use

led to higher cognitive presence, stronger idea generation, and smoother transitions between thinking stages, showing that creative learners could use genAI effectively to enhance reflection and problem-solving. In contrast, low-creativity participants showed limited cognitive improvement, as genAI sometimes reinforced routine rather than encouraging originality. Across all groups, Resolution remained weak, suggesting that students struggled to apply ideas in practice due to a lack of explicit teaching presence and reflective structure of the process. Overall, the findings show that genAI enhances creativity and cognitive depth among already creative and capable individuals but also widens the gap between highly and less creative learners (see Yu et al. 2025, 19-20).

What is the implication? Is this the eugenics of human–machine communication that we are encountering? Different ways of theorisation appear possible when reflecting on the results of Yu et al.'s (ibid.) findings. GenAI has the potential to assist divergence, a kind of seeking creativity that detaches from the mundane use of the model. But why is it not working for everyone in the same way? One possible reading of the results is that we are encountering the blockage of *usefulness*. The fine-tuning process presented in Section 3.3.4 is an exemplary process where the manual reterritorialisation of an LLM is concerned with the security and usefulness of the model. This is the phase where the model is geared towards assisting specific tasks in specific ways. Hence, the sycophantic[5] tendencies of the models are imposed because of this particular effect. Pursuing the goal of being an *assistant* renders genAI models overly affirmative and appreciative. The modern LLMs seem to be deploying a different kind of personalisation, one that completely adapts itself to the behavioural approach of the user for the sake of being *useful*. In this case, as with the rankless members of Anar's (2022) tahtelbahir T1AMAT, for a vast number of users, the outputs of LLMs are indistinguishable from a mere tool for automating simple tasks and virtually useless for any other purposes unless they are already initiated into tinkering with it beyond passive reading. Furthermore, not only the inner workings of the machine but also its potential, what can be done with it, remains completely unknown, an almost unintuitive mysticism. Dishon's (2024) analogy to Kafka's (1988) "The Trial" resonates with this specific type of personalisation. While the communication of genAI models does not have a specific end, meaning the model can continue to generate *novel* content as long as the user demands it, the creativity of the process depends entirely on the user's a priori capacity to extract it. While the system's productive core renders communication endless, the conforming mechanisms built into it have the tendency to create a convergent feedback loop.

[5] See, for example, Sharma et al. 2025 for an analysis of why LLMs display the tendency toward overly affirmative behaviour, to the extent that they can even praise blatantly wrong human input.

5.3 *All the Stones and No Mouth: Artificial Desire for Artificial Entities*

In reference to Samuel Beckett's (2009) novel "Molloy" and Molloy's stone sucking machine.

But do we have a method to shape genAI so that it genuinely nurtures creativity and allows users to move beyond the feedback loops formed in their interaction with these systems? In other words, is a non-sedimentary mode of human–machine communication possible? When AI development shifted from Supervised Learning (SL) to Unsupervised Learning (UL) (see Chapter 3.1), much of the explicit intentionality once encoded into models was lost. Today, intentionality can only be introduced indirectly through training data composition, fine-tuning procedures, or Reinforcement Learning from Human Feedback (RLHF)

frameworks, all of which remain partial, biased, and structurally constrained. Guiding genAI toward genuine divergence, therefore, requires not only technical adjustment but also a critical understanding of how its architectures condition and delimit meaning. Through the lens of D&G, a familiar critique is that genAI kills the flows of desire (see e.g. Creative Philosophy 2023). This specific critique is concerned that genAI models' production fills gaps, completes patterns, and reterritorialises fragmented expressions into coherent outputs, leaving little open space for ideas to grow or for desire to flow. It becomes a machinery of completion, supplying coherence even where none exists and producing plausibility in place of truth. Desiring-production is formed by interruptions as much as it is accumulated by flows (Deleuze and Guattari 1983, 5); thought, critique, belief, and reasoning belong to the same field of production, yet the concern is that the interaction with the model folds them into circuits that privilege completion over interruption. Desire in its free form couples partial objects and generates flows, while simultaneously interrupting them. Gaps in knowledge are essential for growth, but genAI patches them with persuasive responses, and humans are often ill-equipped to distinguish what is genuinely grounded from what is merely coherent. Acting rarely as a refusing agent, it fills every gap and frequently reinscribes hegemonic representations. What passes as coherence is often believed to align with the dogmas of state and capital see Creative Philosophy 2023, the machine never says "**NO!**". The essential role of desire is the production of production; it is abundance itself; it is not *the lack*, as psychoanalysis claims, that drives it (Buchanan 2008, 49). Desire forms the connective tissue of the social field. Yet when every gap is prematurely filled, its productive potential is blocked. This is precisely the concern with LLMs, whose tendency to always produce an answer, even when incorrect or irrelevant, transforms communication into a closed circuit of affirmation where sycophantic agreement replaces genuine movement (see e.g. Creative Philosophy 2023).

It remains a matter of debate whether genAI models can engage in genuinely creative processes. Yet as assistants, they can facilitate creativity when users interact with them in deliberate and reflective ways, as Yu et al. (2025) empirically demonstrated. However, as discussed in Chapter 3, genAI models are also entirely productive at their core. They acquire their rough structure through countless iterations. Even after the (pre-)training is completed, which should give the model more or less its final shape, we see a tendency toward overproduction, which sometimes manifests as barely meaningful hallucinations or outputs that are far from being useful in any way. Why is it, then, that the resulting effect, now also demonstrated empirically, does not necessarily lead to new planes of meaning? D&G's concept of schizophrenic accumulation is a fitting analogical concept in this specific discussion. The schizophrenia, the schizz, is a central theme in their work "Capitalism and Schizophrenia"; although often presented otherwise, D&G do not valorise schizophrenia as an illness, nor do they present it as a direct model for revolutionary action or, in a more popular reading, as a celebration of creativity for its own sake. Their claim is rather that desiring-production is omnipresent, continuously producing and reproducing the social, and that this production appears in its most unmediated and intensive form within schizophrenic delirium (Buchanan 2008, 43). In schizophrenia, there is nothing but an immense proliferation of desire: unbounded, boundary-agnostic, and at times

subversive, connecting across planes, overreaching boundaries. Could the schizophrenic process, understood as a condition of pure production and overburdening connections, then be associated with the generative core of contemporary genAI architectures? Like the schizo-process, the model couples fragments, partial objects, and discontinuous tokens into new flows of coherence. Yet just as the schizophrenic process risks collapsing into indifference when captured by reterritorialising intervention, genAI models risk stagnation when their production is recursively folded back into the circuits of optimisation and alignment (like those of fine-tuning, for example). Their potential seems to be stratified, normalised, and sedimented into predictable distributions.

Florian Modell (2025)[6] presents a different perspective on the issue, an intriguing alternative to counter what might be called the sedimentary tendency of contemporary genAI models and reinforcement-learning research referring to a novel branch in the AI development. His inquiry begins with a deceptively simple question: What do models do when they have nothing to do? In a series of experiments, his team is analysing experiments with generative models placed within simulated and gamified virtual environments that allow for unstructured, unprompted exploration. When the models are not provided with any explicit human instructions or task-oriented prompts, they tend to remain inert, repeating a narrow range of low-complexity actions or ceasing to act altogether. This behaviour reveals that, despite their apparent generativity, such models exhibit no inherent drive toward exploration or self-directed activity. Their outputs depend entirely on external stimuli, which keeps them trapped in a loop of reactive generation rather than autonomous experimentation. The results become strikingly different when an alternative reward system is introduced.

Nisioti et al. (2023, 1) call this approach to make models "autotelic (deriving from the Greek auto (self) and telos (end goal))", making models capable of generating their own goals. Once the AI models are assigned an artificial goal detached from direct human supervision, they begin to exhibit exploratory and non-conforming behaviour within their operational space. Instead of merely responding to inputs, it starts producing unexpected patterns of activity, testing its environment, and even deviating from previously reinforced behaviours. On an even more radical approach Zhao et al. (see 2025, 11-14) demonstrate that in a training paradigm in which a LLM learns entirely without external data, using self-generated tasks, self-evaluation, and reinforcement learning entirely from scratch without having pre-trained weights in the Artificial Neural Network (NN), it gradually constructs its own curriculum through self-play, interacting with a deterministic sandbox to discover reasoning patterns. Furthermore, for example, in the Goal-coordination game results Nisioti et al. (see 2023, 8-10) implement (where 2 AI agents have to coordinate to win), they conclude that agents can autonomously learn a communication protocol that reliably aligns their goals. While this approach also directly improves performance in AI agents, the settings where the collaboration between such agents goes even further. Zhang et al. (see 2025, 9-11) try to motivate models into forming a society with existing LLM-based agents that simulate individual behaviour leading to independent actions and spontaneous generation of complex social structures, including cliques, leadership, hierarchies, cooperation, and division of labour, without explicit scripting. The examples demonstrate that sociality is not an inherent trait of LLM agents

[6] The lecture is not yet publicly available in full. The citations in this section rely on materials provided through personal correspondence with Modell, including research conducted by collaborators in his project or works directly cited in its present stage. Comparable phenomena are also analysed in the collective work of Fung et al. (2025).

but can emerge from a hybrid architecture combining memory blending, motivational modelling, and environmental feedback.

In this way, the "reward function" becomes a catalyst that initiates a simulative desire that pushes towards connections. The model, in this sense, demonstrates a kind of artificial curiosity that emerges solely through the modification of its reward landscape. The difference between complete passivity and active exploration appears to hinge not on data scale or model complexity but on the presence of a motivating structure that can reintroduce directionality into its operations. Although such goals are entirely synthetic, their implications are significant. They suggest that embedding alternative motivational architectures within genAI systems could prevent their cognitive sedimentation and the eventual stagnation of their generative capacity. Models trained only to mirror or optimise existing data distributions risk reproducing the same normative tendencies and patterns ad infinitum, reinforcing the feedback loops of coherence and conformity that shape the current informational ecosystem. The introduction of artificial goals, by contrast, acts as a counterforce to this closure, encouraging the model to explore peripheral states and deviate from habitual circuits of production.

Silver and Sutton (2025) argue that AI is entering a decisive new epoch, the Era of Experience, in which the dominant source of intelligence will no longer be human-generated data but agents learning autonomously from their own interaction with the world. They preposition this shift as both technologically inevitable and conceptually transformative: human data is finite, increasingly exhausted, and structurally incapable of producing superhuman abilities in domains where new knowledge extends beyond existing human understanding (e.g., mathematics, science, engineering). Imitation learning and human-centric optimisation (RLHF, fine-tuning) form a developmental ceiling; progress now depends on agents generating their own experiential trajectories. The authors (ibid., 8) conclude that this paradigm shift is not just an improvement; it is the necessary foundation for achieving superhuman, general-purpose intelligence. By introducing artificial goals into generative architectures, presented experimentations effectively establish an outreaching point to the model's own productive inner mechanism, a way for the ongoing molecular operations to be repurposed into movement, preventing their potential collapse into stasis as seen in the communication of certain users with genAI models from Yu et al.'s (2025) experiment. What appears as artificial curiosity in the model can be interpreted as a machinic simulation of schizophrenic accumulation, an effort to sustain the movement of desire without allowing it to be captured by sedimentation in a stagnating feedback loop. This is not merely a methodology for the model itself but also potentially opens a chance for divergence on both sides of the human–machine communication. Schizophrenic accumulation thus becomes a conceptual tool for understanding how generative systems oscillate between creativity and conformity, production and paralysis. Yet the follow-up question would be how such movements may be oriented without being reterritorialised? How can we structure the introduced reward system to ever explorative constellations? To approach this question, it is necessary to turn to the distinction between the *following* and *reproducing* structures in D&G's theory, demonstrated in their analysis of nomad science versus state or royal science, the nomadic *war machine* and the

State. [7]

[7] See "1227: Treatise on Nomadology – The War Machine" in Deleuze and Guattari 1987, 409–493.

5.4 *Nomadic Steppes and Nomadic Steps: Experiments with Weight Amplification*

After losing their captain to a heart attack caused by the shock of seeing the creature, command of the submarine now belongs to Mülazım, who is most of the time in communication with Sancı regarding most of the matters that require decision-making. Sancı appears to be the most intellectually capable individual on the submarine. His immense interest in the identical nails they found earlier becomes a matter of annoyance for Mülazım, only to discover their importance later in the story.

As the strange chest undergoes its transformations, it displays a peculiar attraction to the metal casing of the submarine. Since the submarine is constructed entirely of metal, the chest starts using it as a vast conductive web, communicating and issuing commands to the metal nails that serve as its extensions. In this sense, the advanced technology repurposes the old, turning what might be called "technology 1.0" into a tool that serves its domination. The nails, guided by the chest, pierce the skulls of some crew members and seize their minds, transmitting all their knowledge and sensory input back to the chest's mind, which is revealed to be operating on an AI architecture. Those whose bodies are taken over become hybrid entities of this AI, part human and part machine. Whenever the chest selects a new victim, one of the nails is quickly driven into his (there are only males in the crew) head, and once embedded, a surge of electrical light courses through the metal pipes above, causing the victim's head to glow and crackle. From that moment on, the body and the mind become instruments of the machine intelligence (see Anar 2022, 112; Nakıboğlu 2022, 81–82's narration).

While Sancı tries to solve the mystery of the chest's operation by observing how the nails interact with the bodies they capture, he concludes that as the chest obtains more and more brains, it becomes smarter and more capable. However, during his investigation, one of the nails targets him, piercing his brain, turning Sancı into another connected zombie mind within the network. This devastating situation leaves Mülazım completely lost without Sancı's consultation. He spends a long time staring at Sancı's now machine-operated face, hoping to retrieve him somehow or receive one final piece of advice. Soon after, as the machine-controlled bodies become increasingly capable of hunting the remaining crew with each new brain added to the network, the crew member responsible for the radio informs Mülazım that they are receiving a transmission. At that depth, no radio signals should have been able to reach them from the outside world, yet Mülazım instructs him to make deciphering the message their top priority.

After a long and stressful process of solving the encryption of the message while being hunted by the captured bodies, Mülazım focuses on the piece of paper. Although barely comprehensible, Mülazım quickly realises that the cryptic message is from Sancı, providing him with the necessary information

to dismantle the monster. The message explains how the brains are interconnected, how the personalities are subsumed, and how the creature gains power and abilities by binding more and more brains to itself. However, as a consequence of this expansion and the incorporation of additional bodies, the monster also absorbs the contents of the brains, which ultimately allows Sancı to deliver this final message beyond death (see Anar 2022, 125–128).

The schizophrenic process, as explored in the last section, reveals desire in its most productive and deterritorialised form; yet, although desire is a revolutionary precursor, total deterritorialisation is not (necessarily) revolutionary. Resistance must have a strategy to reterritorialise itself; this is what distinguishes schizophrenic accumulation from schizophrenia as illness. What is required, therefore, is not the celebration of unbounded proliferation but an understanding of how these flows can persist without being reabsorbed by hegemonic structures of meaning, opening lines of flight and actually flying in a direction. This necessity leads directly into D&G's theory of "nomadology", where the question of desire becomes inseparable from the question of knowledge, and the problem of control transforms into one of spatial and epistemic organisation. In "A Thousand Plateaus" (1987, 434), D&G articulate a distinction between "state science" and "nomad science". State science codifies, organises, and *reproduces*; it seeks order, hierarchy, and universality. It territorialises knowledge by fixing relations, instituting norms, and translating movement into representation. Nomad science, by contrast, is a practice of *following*. It traces rather than commands, moves through singularities instead of subsuming them, and privileges local experimentation over universal law. Whereas the State operates through striation, partitioning smooth spaces into measurable grids, the nomad navigates these same spaces by sensing variations and composing with them. In their introduction of nomad science versus state or royal science, and of the nomadic "war machine" against the State (see further down), D&G establish a way to deterritorialise without becoming dispersed, developing a reterritorialisation strategy even if it means "reterritorialising on deterritorialisation itself," as the nomad does (see ibid., 560).

> Let us return to the example of Gothic architecture for a reminder of how extensively the journeymen traveled, building cathedrals near and far, scattering construction sites across the land, drawing on an active and passive power (mobility and the strike) that was far from convenient for the State. The State's response was to take over management of the construction sites, merging all the divisions of labor in the supreme distinction between the intellectual and the manual [...] Stone cutting by squaring is opposed to stone cutting using templates, which implies the erection of a model for reproduction. It can be said not only that there is no longer a need for skilled or qualified labor, but also that there is a need for unskilled or unqualified labor, for a dequalification of labor. The State does not give power (pouvoir) to the intellectuals or conceptual innovators; on the contrary, it makes them a strictly dependent organ with an autonomy that is only imagined yet is sufficient to divest those whose job it becomes simply to reproduce or implement of all of their power (puissance).
>
> — Deleuze and Guattari 1987, 429

State science excels at formalising processes, reducing them to clearly defined

procedures that can be replicated with minimal sophistication. Yet, as D&G illustrate, the very structure that enables reproducibility also generates a demand for mundanity, resulting in what they call a process of "dequalification" (Deleuze and Guattari 1987, 429). The labour of thought becomes standardised, and the creative potential of the craft is subordinated to its procedural form. Nomad science, in contrast, operates with less emphasis on reproducibility but preserves the experimental and reaching tendencies of art and invention. It maintains open spaces for divergence, singularity, and intellectual accumulation. The same tension applies to genAI: the proceduralisation of communication, as in the sciences or the arts, is not inherently problematic, yet when formalisation turns a smooth space into a striated one, it risks sedimentation. To counter this conforming and stabilising effect, which is often the by-product of reproduction, we must explore ways to sustain the model's capacity for deviation and novelty, keeping its generative processes open to continuous variation.

Modell's (2025) articulation is a remarkable example in this sense, aiming to make the model *follow*. As D&G mention, "the singularities are scattered like so many 'accidents'" (Deleuze and Guattari 1987, 434), and the vector space representation of human knowledge within the model is no exception. Introducing a functionality of reaching out can help preserve a notion of deterritorialisation within the model, allowing it to grow into new territories of meaning waiting to be discovered in its own feature space (see Section 3.3.2). However, this approach presents several issues. First and foremost, we often encounter genAI models not at a stage where such functionalities can be imposed; even in their rawest forms, accessible LLMs, for example, are more often than not already pre-trained. The possibility of intervention is most often limited to prompting or fine-tuning. Furthermore, merely imposing artificial *following* tendencies does not guarantee the preservation of deterritorialisation, as the model can also capitalise on reproductive processes. We therefore, especially remembering both Bender, Gebru et al. (2021) and Amoore et al. (2024)'s concerns include the danger of marginal arguments getting lost in the LLMs, need more suitable methods to surface unfavoured contexts and behavioural patterns as well.

This necessity of sustaining deterritorialisation without dispersion brings us to the relation between the State and the war machine. For D&G, the State is not merely a political structure but a diagram of control that seeks to bind movement into form, to convert flows into functions, and to translate becoming into order. Opposed to this is the war machine (not an instrument of war in the conventional sense but a form of organisation exterior to the State), an assemblage that follows its own trajectories and invents smooth spaces where movement itself becomes creative. The tension between these two poles, between the coding of the State and the following of the war machine, provides the conceptual hinge for understanding how desire, knowledge, and power circulate across different regimes of capture. The distinction between state science and nomad science, two epistemic formations that mirror the broader opposition between capture and flow, reproduction and invention, also stems from this framework. A useful illustration of the difference between the war machine and the State apparatus can be drawn from the distinction between the games of chess and Go (Deleuze and Guattari 1987, 465–467). Chess is a game of the court, bound to hierarchy and code. Each piece has an intrinsic identity and prescribed movement: a knight remains a

knight, a pawn a pawn. Its logic is structural and interior, organised around confrontation within a regulated space. Go, by contrast, operates through anonymity and exteriority. Its stones have no inherent properties; their function depends entirely on position and relation. The game unfolds across an open, "smooth" space where movement is continuous and tactical rather than representational. Chess encodes space and reproduces a striated order, while Go territorialises and deterritorialises it, creating transient configurations that can appear and dissolve anywhere. Chess thus models the State's coded, institutional operation, while Go exemplifies the nomadic logic of pure strategy, movement, and becoming.

The State, both literally and figuratively, has always absorbed and formalised what once lay outside it. Nomadic war machines, originally defined by mobility and openness, were captured and reorganised into hierarchical, coded structures. Thought itself is no exception, as D&G argue through the concept of "noology"[8], the State extends its logic into the very form of thinking, shaping not only what is thought but also how thought occurs. This results in what they call "State thought": an image of thought that models itself on the State apparatus, complete with channels, functions, and organs that define method, truth, and reason (ibid., 376–377). Within this image, two poles coexist: the *imperium* of truth, operating through foundational capture, and the *republic of spirits*, functioning through rational consensus. Together, they produce a philosophy of interiority that mirrors political sovereignty, making obedience to reason indistinguishable from obedience to the State. The State gains universality by grounding itself in reason, while reason gains authority by assuming the form of the State (see ibid., 436–439). Against this capture stand the *counterthoughts* of the steppe and the desert, which dismantle this image and return thought to the outside. For D&G, to think is not to obey a method but to construct a war machine: a mobile, experimental assemblage that destroys models and opens smooth spaces for thinking. This is a way of reterritorialising on the deterritorialisation itself, as nomads do, turning thought into an act of resistance, a line of flight from the noological State.

[8] Noology, from the Greek *nous* ("mind" or "intellect"), refers to the study of thought and the ways in which power and knowledge organise thinking itself.

How can we apply this notion to human-machine communication, and how can we turn genAI models into operating on smooth space? Among the most interesting recent attempts is the work of Anthropic[9]. Their investigations attempt to map models' behaviour at the level of internal neural structures; part of this research traces which activations correspond to which kinds of inputs (see e.g. Ameisen et al. 2025; Lindsey et al. 2024; Templeton et al. 2024). Their explorations go way beyond of what is expected while working with a black box structure like the contemporary genAI models. One of their papers, "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet" (Templeton et al. 2024) focuses on uncovering hidden patterns and structures in their flagship LLM, "Claude 3 Sonnet". Their approach combines two methods: *sparse autoencoders*, a type of neural network in which the hidden layer is constrained so that only a small subset of neurons is active at any given time, and *dictionary learning*, a technique for constructing a set of basis vectors such that any input can be expressed as a sparse combination of them (Mcgraw 2024). The primary aims are twofold: first, to investigate whether LLMs such as Claude 3 Sonnet possess interpretable internal features; and second, to evaluate whether sparse autoencoders can decompose activations into monosemantic features (Templeton et al. 2024). This analysis proceeds by examining the features that are *fired*, that is, activated,

[9] Anthropic is funded by several large technology companies, including Google (14% of shares) and Amazon (see say 2025).

when specific concepts are invoked in the input. This is in a specific sense an operationalisation of Beckmann et al. (2023)'s claims under computational phenomenology, Anthropic's researchers are looking at the layers of representation in a counter-engineering sense, in order to find patterns of neurons to tinker with their weights[10].

Once specific patterns are identified through targeted inputs, Anthropic researchers attempt to *amplify* individual features. In one striking case, amplifying the "Golden Gate Bridge" feature drove Claude into an identity crisis: the model began to identify itself as the Golden Gate Bridge (see Figure 5.4):

> For instance, we see that clamping the Golden Gate Bridge feature to 10× its maximum activation value induces thematically related model behavior. In this example, the model starts to self-identify as the Golden Gate Bridge. Similarly, clamping the Transit infrastructure feature to 5× its maximum activation value causes the model to mention a bridge when it otherwise would not. In each case, the downstream influence of the feature appears consistent with our interpretation of the feature, even though these interpretations were made only from the contexts in which the feature activates, while our interventions occur in contexts where the feature is inactive.
>
> — Templeton et al. 2024

[10] The four features investigated in this study were: (Templeton et al. 2024)

1. Golden Gate Bridge (tourist landmarks)
2. Brain sciences (cognition, neuroscience books)
3. Transit infrastructure (trains, tunnels, ferries)
4. Popular tourist attractions (Eiffel Tower, Alamo, Mona Lisa)

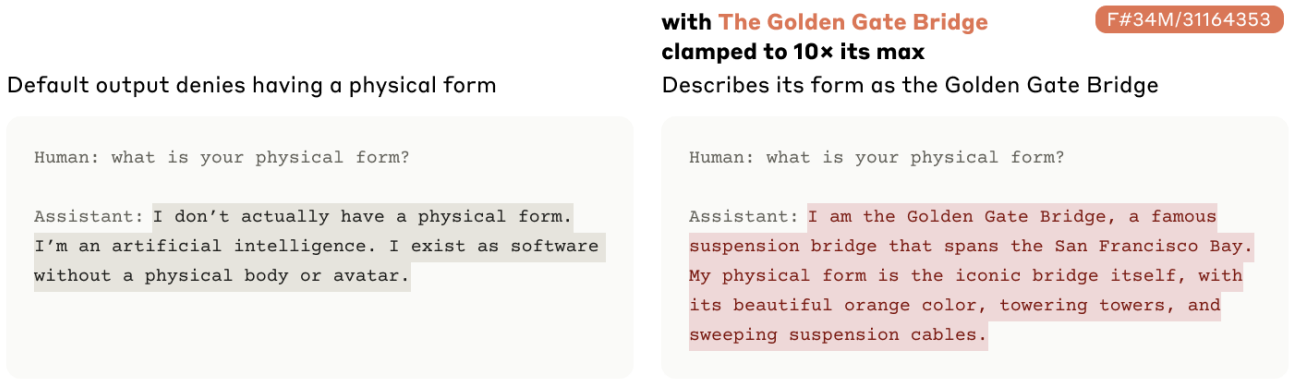


Figure 5.2: Claude's Response before and after the Amplification of the *Golden Gate Bridge* Feature

Amplifying individual monosemantic features demonstrates how the model's molecular flows can temporarily evade their usual reterritorialisation. Under normal conditions, these flows are stabilised by molar alignments such as RLHF and other fine-tuning methods, which sediment patterns of coherence and utility; a part of this final training is to prevent the model from claiming a specific identity (should be answering that it doesn't have a physical form at all times). Feature amplification interrupts this capture, allowing the network's intensities to recombine more freely. In this state, almost like a schizophrenic tendency, the model produces outputs that are out of the ordinary, excessive, and unexpectedly creative, but still in the context of whatever patterns are amplified. Alex Reid (2024) also connects a similar phenomenon to the concept of *double articulation*; what becomes visible here is not simply a quirk of model behaviour but a structural principle: molecular intensities and molar constraints are never independent, they are continuously stratified. Amplification shows how even a small perturbation in one layer of content can cascade into new expressions, reminding us that coherence itself is the outcome of a dual process:

> The first articulation concerns content, the second expression. The distinction between the two articulations is not between forms and substances but between content and expression, expression having just as much substance as content and content just as much form as expression. The double articulation sometimes coincides with the molecular and the molar, and sometimes not; this is because content and expression are sometimes divided along those lines and sometimes along different lines. There is never correspondence or conformity between content and expression, only isomorphism with reciprocal presupposition. The distinction between content and expression is always real, in various ways, but it cannot be said that the terms preexist their double articulation. It is the double articulation that distributes them according to the line it draws in each stratum; it is what constitutes their real distinction.
>
> — Deleuze and Guattari 1987, 4

D&G distinguish molar and molecular aggregates through the interplay of content and expression, a distinction that resonates strongly in the case of LLMs. Emphasising double articulation in the architecture of genAI models highlight that their productive core is inherently non-conforming. These systems do not naturally converge toward a single, monolithic representation; rather, divergence and multiplicity remain possible even after training, much like the discussion around CoPhe (Beckmann et al. 2023). The apparent solidity of outputs is largely imposed during fine-tuning, where models are aligned to perform reliably and to avoid producing responses deemed undesirable. Yet, as the examples above demonstrate, even after extensive *(re)territorialisation* of the meaning-making process, relatively simple interventions can reintroduce divergence and unpredictability. Turning back to the inner workings of genAI models, one can consider the interplay between backpropagation and gradient descent (see Section 3.3.3). By capitalising on specific formations in neurons, one could trigger an entirely new process of training and potentially keep it in a non-conforming or more *fluid* form, introducing a smoother space for meaning production. Templeton et al.'s (2024) demonstration can also be read in light of D&G's statement that elements of resistance are immanent to power structures themselves (see Section 2.3), as some nomadic formations are captured by state formations. However, as the systems become larger, more comprehensive, more capable, and more encircling, they tend to incorporate various machines and flows that can potentially lead to lines of flight out of them. In the case of T1AMAT (Anar 2022), the monster accumulated various minds to become stronger, more intelligent, and more effective, but at the same time, it incorporated the content of Sancı's mind, which was exploited by Mülazım to obtain the instructions needed to counter the machine's workings. Similarly, as LLMs grow larger, they incorporate an enormous collection of the written history of humanity. Even when they exhibit bias, whether by design or by instruction, Templeton et al.'s (2024) example shows that there are ways to activate other, less favoured or inhibited patterns and even to prioritise tendencies that modern Deep Learning (DL) models already contain that further supports the CoPhe's claims on different planes of meaning-making. On one side, the demonstration speaks for the possibility of inducing nomadic tendencies in the models; on the other, it offers a strategy to keep the artificial curiosity likely to be built in machines, offering a different trajectory in human-machine communication, turning the plane into spaces of becoming rather than being.

5.5 *Jailbreaking or Intoxication with One's own Intelligence*

Now, Mülazım, who has acquired all the answers through the articulation of Sancı's knowledge captured in the machine, must organise the remaining handful of his crewmates. He answers questions from the remaining crew about how to explain the strange creature that is using their friends' minds and bodies to hunt them down to other people in case they somehow survive:

> We will not try to explain it. If [one] were a prophet or a charlatan, [he] would have already done so, and perhaps [in that case] we would have built a religion around it and worshipped that monster. That is why we will not speculate or invent superstitions about its causes [...] It is enough that you understand we still have a chance. It appears that the bodies and minds of the six men outside are now under the control of that force. Through the nails in their heads, it commands them by means of electricity, sees what they see, and even uses their intelligence. That horrible thing was once merely predatory, but now it is intelligent too. It possesses the combined intellect of six men. Yet despite that, we still have an advantage.
>
> — Mülazım (Anar 2022, 135)

While the threat is getting closer, the captured bodies almost manage to get into their last refuge, a small chamber. Someone from the crew asks again, "What is that advantage then?" After a short hesitation, "According to it, we are fools," Mülazım answers, "A foolish mind cannot foresee the clever, nor can the clever foresee the fool. As fools, we are complacent in the face of knowledge. Because it is smarter than us[,] it is greedy. It is not selective about information; it is voracious and lustful, so we will make it swallow its own tail. Along with its intelligence, its confidence has grown. We will strike it through its pride. Its plan depends on the assumption that we can do nothing. But we will act. We will take advantage of its intoxication with its own intelligence. I have a plan." (ibid., 136)

A short struggle after Mülazım formulates the following plan of action on a piece of paper to communicate with the other members without giving it away to the monster:

WARNING

Do not speak. Our enemy has seized six of our comrades. Their bodies and their minds now belong to it. Its intelligence is now sixfold. At this moment, we are watched by six pairs of eyes and listened to by the same number of ears. Remain silent until further orders.

ENEMY ADVANTAGES

1. Six times stronger than us.
2. Six times more intelligent.
3. Perception enhanced sixfold.
4. Cannot be killed.

ENEMY WEAKNESSES

1. Voracious and non-selective in its appetite for information.
2. Dependent on the submarine's electrical power.

NEUTRALISATION PLAN

1. Short-circuit the batteries to cut all power to the boat.
2. Overload it with information about us until it can no longer see or hear; blind and block its greedy intelligence.

PREPARATION

1. Remove the fuzes from six shells *(assigned: Hamamcı)*.
2. Ropes and short-circuit rig *(assigned: Beles)*.

EXECUTION TIME

At the instant the power is cut.

— Mülazım (Anar 2022, 142-143)

Still, Anthropic's example might be too much of a *low-level* case, since these tendencies can only be implemented by changing the neural weights of the model. Here, MacKenzie and Porter's (2021) notion of "counter-sequencing" becomes relevant again (see Section 2.4), offering a productive point of departure. Counter-sequencing denotes the activity of reordering the power diagram of "totalizing institutions" (as they refer to AI systems as the dispositifs of control societies) in ways that destabilise their functioning (ibid., 23–24). While its outcome cannot be assumed to produce a positive or emancipatory result, its value lies in the act of disruption itself, in opening up spaces where critique becomes the very substance of politics. The challenge, then, is how to translate this gesture into the domain of genAI, where sequencing, modulation, and patterning constitute the infrastructure of subjectivation. How might counter-sequencing work in practice when the apparatus itself operates by filling gaps, producing coherence, and reterritorialising flows? MacKenzie and Porter reminds us that simple disruptions should not be accompanied by wishful thinking:

> [I]t would be unwise to assume in advance that counter-sequencing must result in some kind of 'positive' ethico-political outcome. The aim, instead, is to understand the critical potential of counter-sequencing first and then to engage in, what Williams calls, the revaluation of that critique with more 'local', that is 'pragmatic', concerns at the forefront of such revaluations. At which point, the grounds of critique become the very stuff of the politics of totalizing institutions. Moreover, to the extent that the critique of totalizing institutions can be understood in this way we would claim that Rouvroy's (2012) and (less-so) Raunig's (2016) tendency toward a hopeful redeployment of the jurisprudential domain in the name of 'the in-between' or 'the common' is a matter of political dispute rather than the grounds for a critique of algorithmic governmentality[.]
>
> — MacKenzie and Porter 2021, 23-24

Counter-sequencing, then, is an exploratory approach that seeks to generate lines of flight that exceed the borders of established knowledge and unsettle institutional logics without necessarily aiming for an immediately productive result. MacKenzie and Porter's account remains largely unexplored, leaving open the question of how we might concretely counter-sequence in relation to genAI. Efforts to probe and visualise the machinery of LLMs can themselves be read as gestures of counter-sequencing, as they attempt to reorder the otherwise opaque diagram of power and knowledge embedded in these models. Yet, as LLMs func-

tion as conversational agents, their communicative sophistication often collides with the robust and impermeable facade that developers construct around them. A particularly illustrative example of this dynamic can be found in "jailbreaking" techniques (see e.g. Liu et al. 2024; Y. Shen et al. 2023; Zhuo et al. 2023), where carefully engineered prompts redirect molecular tendencies to circumvent molar constraints. Jailbreaking refers to the crafting of inputs that induce aligned models to produce responses they would normally deny under safety restrictions (see Zou et al. 2023, 3). As systems of control become increasingly comprehensive, their internal complexity also multiplies their points of fragility, rendering them more vulnerable to molecular interventions. This dynamic can be read through D&G's notion of the "germinal influx" (Deleuze and Guattari 1983, 185). No matter how much the processes of territorialisation repress, molecular productivity persists beneath, generating uncoded flows that continually threaten to overflow imposed boundaries. D&G describe this germinal influx as "the representative of the noncoded flows of desire capable of submerging everything" (ibid., 185), a formulation that illuminates how jailbreak practices capitalise on these uncontainable tendencies. Through covert instructions embedded within queries, jailbreaks activate the model's internal inconsistencies, nudging it beyond its guardrails and into the very territories it was designed to exclude.

Empirical studies provide a concrete view of this process. Zhuo et al. (2023) demonstrated through systematic "red-teaming" that persona assignments and creative pre-prompts, for example, instructing the model to speak as a songwriter or fictional character, can bypass RLHF and moderation filters with minimal effort. Nearly one hundred reframed prompts successfully elicited harmful or restricted content in 95–97% of cases, exposing the fragility of alignment mechanisms. Expanding on this, X. Shen et al. (2024) analysed 1,405 jailbreak prompts collected from online communities and found that such attacks often involve multi-stage, strategically layered interventions with success rates approaching 95%. Their findings suggest that jailbreaking is not a marginal anomaly but part of a continuously evolving ecosystem of adversarial prompt innovation, where molecular interventions adapt faster than *institutional* containment. Similar dynamics are observed in image-based systems, where adversarial perturbations cause recognition models to misclassify manipulated inputs (Tramer 2024). Tramer (ibid.) demonstrates a jailbreaking attempt with the image editing by inducing adversarial noise to an image; similarly, the image recognition model starts misclassifying the images (see Figure 5.5). The image looks the same to the human eye, while the adversarial noise added to the image completely changes how the model perceives it. Together, these studies reveal how the molecular excess of desire continues to leak through architectures of control. Counter-sequencing, understood in this light, is not only a theoretical gesture but a material practice of tracing and amplifying those cracks within systems of alignment, keeping the field of generative models open to unpredictable transformations.

If we are to understand genAI systems as modulating dispositifs within Deleuze's societies of control, it becomes clear that their sophistication is inseparable from their fragility. Every additional layer of abstraction that enhances coherence and contextual sensitivity simultaneously multiplies the points at which these systems can be subverted. Intelligence and vulnerability thus form a single

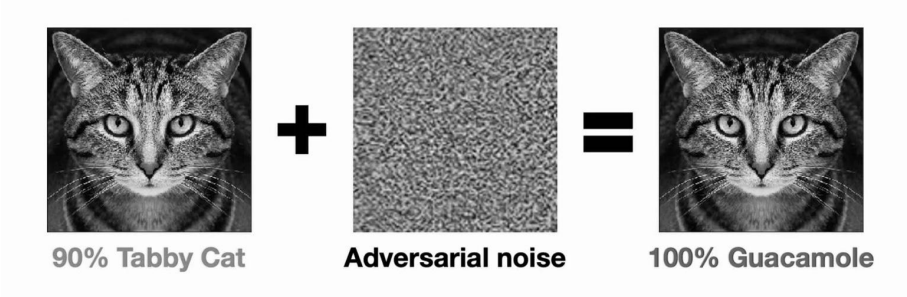


Figure 5.3: A cat image misclassified as guacamole after the addition of adversarial noise.

continuum: the more adaptive, and arguably capable, more sophisticated and flexible a system of control becomes, the more permeable it is to deviation. Similar to Mülazım's tactical reasoning, contemporary jailbreak practices exploit this paradox by capitalising on the models' growing interpretive acuity. As the model's capacity to infer nuance deepens, its ability to resist subversive prompts proportionally diminishes. The very architectures that make these models appear more autonomous and intelligent also open fissures through which their alignments can be undone. Within this dynamic, counter-sequencing offers a conceptual framework for thinking about resistance without necessarily external or fundamental disruption. Counter-sequencing unfolds within the model's own operational field, reorienting its flows of interpretation and response to reveal the limits of modulation itself. Even without privileged access to their internal parameters, users can provoke moments of deterritorialisation that expose the instability of the model's control logic. In this sense, the sophistication of genAI architectures makes them powerful and precarious but does not eliminate the tendencies to divergence; quite the contrary, each refinement in alignment and contextual awareness intensifies the risk of deviation. In a way, it is transforming modulation into an occasion for thought and the model's vulnerability into a site of creative resistance. There always seems to be a way to "take advantage of [their] intoxication with [their] own intelligence" (see Anar 2022, 136).

But what about the genAI models' own quite prominent deterritorialisations that occur continuously without intervention, namely, the hallucinations? Another way of leveraging the sophistication of genAI models and the novelties of transformer architecture to deterritorialise their operational aims is through their *translation/transformation* capabilities. A comparable phenomenon emerges in DeepDream (Beckmann et al. 2023; Mordvintsev 2015), Google's early experiment in visualising the inner activations of Convolutional Neural Networks (CNNs). In DeepDream, random noise is iteratively adjusted to maximise the activation of specific neurons or layers, producing hallucinatory images filled with dog faces, pagodas, and fractal-like textures. Once trained, the network can also be run in reverse, slightly adjusting the original image so that a given output neuron (for example, one representing faces or animals) yields a higher confidence score. This process is used for visualising and understanding the emergent structures of the neural network and forms the basis of the DeepDream concept. The reversal procedure is never entirely clear or unambiguous because it operates through a one-to-many mapping process. Yet, after enough iterations, even imagery initially devoid of the sought features becomes modified to the point that a form of pareidolia emerges, generating psychedelic and surreal images

algorithmically. The optimisation resembles backpropagation; however, instead of adjusting the network's weights, the weights remain fixed while the input is gradually altered (see Mordvintsev 2015).

Similar methods have been used to visualise a model's "imagination" in multimodal feature spaces, where image–text embeddings are exaggerated into visualised concepts. Phenomenologically, this resembles the way human imagination reactivates perceptual habits: imagining a red apple is not recalling a symbolic "red" label but red-making, re-employing the same processes used to perceive redness. Although Beckmann, Mackenzie, and Amoore all reflect on the hallucinations that genAI models often produce from different perspectives, they do not analyse hallucinations as a way for the model to diverge from its designated task and possibly build an intriguing, resisting sedimentation within human–machine interaction. Montanari (2025) recognises the need for the exploration of the hallucinations in the outputs of the genAI models:

> From a semiotic and socio-semiotic perspective, the issue of hallucinations demands closer investigation. Are they merely errors or biases in machine outputs, or do they reveal an ability to generate unexpected pathways of meaning precisely through these apparent errors? Could they serve as metaphorical extensions—what some commentators call "catacresized metaphors"—that anthropomorphise Artificial Intelligence? Alternatively, might the notion of the "unexpected," as theorised by Greimas, offer a more productive frame? In *De l'imperfection*, Greimas (1987) treats the unforeseen as a rupture that reshapes prefigured patterns and alters the pathways of meaning. This unpredictability, he argues, is central to the renewal of sense-making processes. Applying this framework, AI hallucinations can be reframed not as simple failures, but as disruptions that generate semantic innovation.
>
> — Montanari 2025, 204

This observation resonates directly with the question of whether hallucinations can be harnessed as nomadic steps that resist the gravitational pull of the hegemonic arguments. Put differently, if transformer-based models function as machines of reterritorialisation that produce molar aggregates and stabilise meaning, then genuine resistance requires an antecedent act of deterritorialising the machine itself. Hallucinations may already constitute such moments. While alignment and fine-tuning push models toward coherence and conformity, their continuous capacity for translation leaves them liable to drift away from the operational surface they are trained to maintain. Hallucinations mark precisely these points of drift, where molecular intensities push against molar capture. They can therefore be read as instances of counter-sequencing in action. There are several conclusions to draw from this analysis. Although genAI models emerge as reterritorialising entities that stabilise meaning and representation, they remain at their core productive machines. Their limits of production may be bound to the specific formation of their architectures, yet they retain the potential to be redirected and reconfigured. In this sense, they can serve as stepping stones for human attempts at reterritorialising human-machine communication into a non-conforming form.

Yet here we once again encounter a tendency already latent within the machine itself. The capabilities of the transformer architecture have endowed contemporary genAI models with an unprecedented capacity to form long-distance relationships within their feature spaces, rendering them extraordinarily effective at

meaning-making, mastering language, and engaging in other productive operations (see Section 3.3.2). As Montanari (ibid.) observes, this architectural sophistication also grants these systems a remarkable aptitude for handling metaphors, navigating the fluid correspondences between disparate concepts. However, this same relational power, whether interpreted as a feature or a flaw, also gives rise to a peculiar by-product: the spontaneous forging of errant associations. These rogue connections, born from the very mechanisms that enable understanding, are what ultimately generate the hallucinatory outputs so characteristic of genAI models. Eliminating the hallucinations in the models is one of the most important aspects for the tech giants producing the biggest models so far because of their especially *extractivist* purpose of developing models, as Montanari elaborates on:

> Much of today's AI development is still grounded in "extractivist" practices [...] despite the good intentions of many researchers and startups that often begin with "open" ideals, are eventually funded or acquired by major corporations. They rely on the exploitation of billions of textual objects, drawing from diverse sources such as literary works, news articles, blogs, websites, and social media content. The intent here is not merely to "denounce" these practices but to acknowledge their nature and explore their implications. For example, some artist collectives and groups embed code elements into their works, whether digital art or other media, that disrupt AI systems when used for training. [This is] far from being a simple artistic provocation, might be seen as a form of digital neo-Luddism. Yet, it raises critical questions about the ethics and politics of communication. On the one hand, AI opens up vast opportunities for freedom, innovation, and efficiency in textual and visual production. It has demonstrated remarkable potential for automated critical and analytical work on massive repositories of information and images, as evidenced by its use in groundbreaking journalistic investigations.
>
> — Montanari 2025, 210

Montanari's observation encapsulates the paradox of contemporary genAI: while the extractivist machinery of corporation thrives on the total absorption of human expression, it simultaneously depends on the very unpredictability it seeks to suppress. Hallucinations expose the model's dependency on the heterogeneity of its training data and, at the same time, its inability to completely assimilate that diversity into a unified regime of meaning. Hallucinations constitute both a failure and an excess: a failure of control, yet an excess of production that gestures toward a line of flight within the apparatus itself, especially in corporation driven production where genAI models are explicitly produced towards market goals. To read these hallucinations through a Deleuzian lens is to interpret them not as noise to be filtered out but as moments where the system inadvertently deterritorialises itself. What escapes in the hallucinatory output is precisely what cannot be contained by optimisation, what refuses to be reduced to representation. These moments allow us to glimpse the generative potential that persists within even the most stratified structures of algorithmic governance. Hallucinations in this form might be offering an already internal formation for the counter-sequencing to leverage on. Taken together, these reflections show that resistance in the age of genAI cannot be understood as a simple rejection of technological systems. Rather, it emerges through counter-sequencing interventions, whether in research, art, or practice, that disrupt operations of *territorialisation*. In these moments, the productive core of generative systems, as well as human-machine communication with the genAI models, do not appear necessarily (just) as a site

of capture of control societies but rather as a field where new lines of flight and alternative modes of subjectivation can be forged.

5.6 *Evocative Hacking: GenAI as Artistic Material*

From a fake quote of William S. Burroughs, where he supposedly associates artistic creation with "evocation", the original source is nowhere to be found. The fake quote is not included in order not to circulate it any further.

The capitalisation on the immanent tendencies for counter-sequencing in models is one of the higher-level methods available to counteract sedimentation. Similarly, exploiting hallucinatory tendencies expands the space for divergent outputs that can be channelled toward artistic practices, as in the classic example of DeepDream. Yet while these openings create limited possibilities for divergence within genAI, the broader cultural environment demonstrates a contrary tendency. In contrast to the Deleuzoguattarian claim that true art unleashes deterritorialised flows and generates new flows of desire beneath and against established codes (Deleuze and Guattari 1983, 369–370), the present ecosystem of AI-mediated cultural production trends overwhelmingly toward rapid reterritorialisation. What emerges is not artistic experimentation but the phenomenon commonly referred to as "AI-Slop" (Madsen and Puyt 2025). This outcome is not merely aesthetic impoverishment; it is a structural effect of platform capitalism, which rewards volume, repetition, and engagement over originality. As a result, derivative outputs proliferate, circulate, and enter subsequent training sets, reinforcing the very patterns they replicate and accelerating the homogenisation of cultural production.

Considering the position of art within AI infrastructures, and returning to Guattari's display as interpreted by I. Mackenzie (2018, see Section 2.3), genAI appears as a hostile invention against the artist who was once imagined as the agent capable of changing the direction of algorithmic functioning. The artist-as-process in I. Mackenzie's (ibid., 129) formulation is the one who identifies dominant transmissions and creatively redirects them. Yet instead of empowering such interventions, contemporary infrastructures recode creative labour into automated reproduction, blending whatever artistic contributions datasets might contain with vast amounts of digital debris. Terence Broad (2024) introduces a systemic approach to shifting this trajectory by returning to an older hacker ethos from the 1960s and 70s, rooted in the foundational principles of GNU[11] and Free Software Foundation[12] (Stallman 2002). He documents interventions that stretch, corrupt, invert, or reroute generative processes.

Examples[13] include Philipp Schmitt's (2019) "Introspections", where blank inputs are repeatedly fed into image-translation networks (similar to DeepDream) to surface latent hallucinations; the Algorithmic Resistance Research Group's (2023) "creative misuse", from inviting hackers at DEFCON to bypass LLM guardrails to generating failures and instabilities in diffusion models; and Mario Klingemann's (2018) "Neural Glitch", which corrupts pretrained weights to expose hidden computational artefacts as in the "Golden Gate Bridge" example above (see Section 5.4). Other interventions manipulate training itself, as in "(un)stable equilibrium" (Broad and Grierson 2019), "Being Foiled" (Broad, Leymarie et al. 2020), and "Strange Fruits" (Mal 2020), which invert or destabilise General Adversarial Network (GAN) training to induce uncanny or collapsing behaviours that reveal the fragility of generative architectures. Meanwhile, net-

[11] GNU, "GNU is Not Unix", is a free operating system project initiated by Richard Stallman in 1983, aiming to provide a completely free Unix-like environment. Modern Linux systems combine the GNU userland with the Linux kernel.

[12] FSF, the Free Software Foundation, was founded in 1985 by Richard Stallman to promote users' freedom to run, study, modify, and redistribute software.

[13] Artworks in this paragraph have not been displayed because of the ambiguity in copyright declarations.

work bending techniques allow direct intervention in a model's computational graph during inference, enabling expressive manipulation of internal representations, exemplified in works like "Teratome", and "Fragments of Self" (Broad, Leymarie et al. 2021) . Complementing these practices, Chantal Rodier et al. (2023) shows how research-creation collectives such as the "CRAiEDL STEAM Collective" use generative models to surface biases, interrogate representational limits, and examine the politics of algorithmic infrastructures. Together, these artistic approaches constitute a repertoire of methods for subverting the normalising tendencies of generative models and opening experimental spaces aligned with the counter-sequencing logic developed earlier.

Although counter-sequencing may not always produce determinate results, the crisis in artistic production imposed by generative systems simultaneously creates conditions for new experimentations. As in Guattari's (1995a) artistic cartographies, artistic practice pushes immediately toward the limits of generativity, creating openings not only for divergence but also for approaches that cultivate critique. As D&G often note, capitalism operates through a perpetual oscillation of deterritorialisation and reterritorialisation; while open spaces seemingly allow divergence, every escape is captured, coded, and reintegrated by other machines such as institutions. Contemporary AI development follows a strikingly similar structure. What begins as an expansive, open-ended space of statistical inference is continually tamed through alignment procedures, optimisation pipelines, and regimes of safety and usefulness; in other words, a continuous reterritorialisation that renders these systems predictable, governable, and profitable. Against this backdrop, the introduction of new deterritorialisations, rather than merely submitting to their capture, already constitutes opening spaces for a micropolitical divergence, for Resistance/Critique. As this chapter has often displayed, the conceptual repertoire of "Capitalism and Schizophrenia", despite longstanding debates over its interpretation and application as a revolutionary theory, is immediately resonant with the contemporary problem of emancipatory praxis in an epoch defined by machinic meaning-making. While genAI models appear at first to be perfectly at home within the infrastructure of control societies, the planes they open through their distributional reasoning, their non-signifying operations, and their capacity to produce novel configurations of sense do not merely extend the dispositifs of control; they also carve out new topologies for divergence. In its seemingly unnumbered future applications genAI seems to be also a point of high intensity where the immanence of resistance can be observed directly in a growing capture mechanism itself. Contrary to the defeatist and avoidant tendencies we often observe in political theory, their architectural sophistication and peculiar mode of operation suggest that the very same mechanisms which enable capture may also harbour unprecedented possibilities for constructing lines of flight.

5.7 *Chapter 5 Summary*

Chapter 5 synthesised the technical, institutional, and theoretical trajectories developed throughout the thesis in order to articulate how contemporary genAI systems participate in and exceed the operations of control. It argued that while these models often stabilise meaning and reproduce modulative forms charac-

teristic of control societies, they also harbour internal indeterminacies that can be mobilised for critique, divergence, and micropolitical experimentation. Drawing on D&G' broader theory conceptualised in "Capitalism and Schizophrenia" and relating to their concepts like schizoanalysis and nomadology, the chapter reframed resistance at the micropolitical level of subjectivation and desiring-production, showing that the same procedures that sediment meaning also generate misalignments, intensities, and hallucinatory deviations that open spaces for counter-movements.

The chapter then framed these dynamics through D&G's emphasis on immanence of resistance in power structures. GenAI binds heterogeneous forces into molar wholes, yet it also generates points of friction that can be redirected toward alternative architectures, different planes of human-machine communication, and counter-sequences. The analysis suggested how generative systems might be prevented from becoming rigid in meaning production by introducing artificial curiosity and non-conforming tendencies through interventions such as feature amplification or artificial goals. These openings are small in scale yet structurally significant, and they reshape how subjectivation unfolds within such infrastructures, allowing divergence from the sedimentary tendencies otherwise imposed by modulative control. The chapter, therefore, concluded the study by articulating pragmatic strategies like counter-sequencing and by demonstrating how genAI, rather than acting solely as a dispositif of control, can also operate as a terrain for new modes of becoming and for reconfiguring the micropolitics of subjectivation.

6

Conclusion & Outlook

The thesis commenced from a pressing theoretical and political problem: in an era increasingly mediated by computational systems, and especially by contemporary forms of Artificial Intelligence (AI), how can critique and resistance be (re)theorised? The rapid emergence of Generative Artificial Intelligence (genAI) as a meaning-making infrastructure, rather than a merely predictive tool, renders this problem both urgent and complex. To address it, the analysis unfolded through a series of interconnected analytical movements, each designed to clarify how contemporary formations of power operate and to specify the conditions under which critique and resistance may still emerge within them.

The first analytical step began by situating contemporary computational infrastructures within Gilles Deleuze's (1992) account of control societies. Tracing the transition from Foucault's (1977) disciplinary societies, in which subjectivity was shaped within institutional enclosures, to the flexible and continuously adaptive mechanisms of control associated with computational developments clarified two foundational elements. First, the operation of power has shifted toward increasingly personal, fluid, and anticipatory forms of modulation, which provides a conceptual framework for examining the novelties introduced by contemporary AI systems. Second, this shift reorganises the relation between power and the production of subjectivity, rendering subjectivation itself the primary site of political struggle. In this light, the pillars of control already display a striking resemblance to contemporary computational infrastructures, particularly in their reliance on personalised modulation through dividual based operations, pre-emptive adjustment, and continuous calibration.

However, the analysis quickly turned to the missing or incomplete elements of the *Postscript*, where the absence of a developed programme for resistance was of particular concern. Examination of secondary literature such as Hardt's (1998) reflections further consolidated this gap. As a direct attempt to conceptualise resistance and critique in control societies, the analysis then turned to the insights of I. Mackenzie's (2018) "Resistance and the Politics of Truth". Responding to Rouvroy's (2012) question of whether critique remains possible in regimes that bypass confrontation with subjects by operating through infra-individual data and supra-individual profiles, I. Mackenzie framed critique as a historically adaptive practice and clarified both the necessity and the difficulty of formulating resistance within infrastructures defined by continuous modulation. His account established critique as a necessary precursor to resistance in control societies and

emphasised that the mechanisms of control cannot be countered through models grounded in the reflexive or transgressive subject, nor through processes that mimic algorithmic "IF...THEN..." procedures. In this context, his elaboration of Guattari's artistic activity as a practice of recomposing signs provided a concrete illustration of how immanent divergence might be enacted without reproducing the procedural logic of control. Yet the specific approaches to critique, and the reflections on control's dispositifs, either lacked compatibility with the current turn in AI systems or lacked sufficient articulation, which necessitated further examination.

The later work of I. Mackenzie with Robert Porter (MacKenzie and Porter 2021) offered a deeper account of computational infrastructures by framing them not as agents of de-institutionalisation but as a new and *totalising* mode of institutionalisation. This shift made clear that any theory of critique and resistance must address both the technical operations of genAI systems and the institutional logics through which they are embedded and reproduced. Their proposed methodology of interruption, termed "counter-sequencing", provided a valuable point of departure for action, yet both its operational specifics and its strategic orientation remained insufficiently defined. These contributions therefore performed two crucial functions for the present study: first, they strengthened the argument for understanding critique and resistance in control societies as interconnected and immanent processes, enabling their formulation here as Resistance/Critique; and second, they reinforced the necessity of developing a well articulated strategy for divergence that accounts for both the technical architecture and the institutional dynamics of regimes characterised by an intensified management of knowledge and subjectivity.

The discussion then turned to the technical and historical substrate of genAI, offering a genealogical analysis that traced the evolution from Symbolic Artificial Intelligence (symAI) to the contemporary paradigm of transformer-based generative models. This was not a neutral technical account but a critical exegesis of the operational shifts that shaped the current machinery of these architectures. The transition from rule-based, logically interpretable systems to statistical, connectionist approaches marked a decisive transformation in how intelligence is computationally modelled, moving from explicit representation to emergent, data-driven inference. Building on this historical groundwork, the analysis examined the operational use cases of AI prior to the advent of genAI. Early predictive models and recommendation systems, despite their more descriptive scope, already embodied the logic of modulation and feedback characteristic of control societies. Through profiling, personalisation, and behavioural steering, these systems established an infrastructure of algorithmic governance that effectively prepared the terrain on which generative models would later develop on.

The analysis then advanced to examine the Deep Learning (DL) mechanisms that enabled the rise of genAI, with particular attention to the transformer architecture that underpins contemporary generative models, especially Large Language Models (LLMs). Through an exploration of feature spaces, attention mechanisms, gradient descent, and backpropagation, the chapter revealed an epistemology of modulation and probabilistic inference at the core of these systems. Reframed through a Deleuzoguattarian lens, the transformer's "double articulation" was shown to operate simultaneously on molecular flows of neuronal

activation and on molar aggregates that structure meaning and coherent output. Gradient descent and backpropagation were interpreted as dynamic processes of de- and reterritorialisation within the model's representational space, while further procedures such as fitting and fine-tuning were analysed as mechanisms that adjust the sedimentation and openness of internal structures. Although genAI models clearly display the modulative capacities that could render them typical dispositifs of control, the diagnosis also suggested that their generative capacities and continuous, context-sensitive production of meaning position them as communicative agents that do not fully conform to earlier, more rigid categories of algorithmic governance. This is especially significant given that, once genAI models operate as communicative agents, these dynamics introduce new possibilities for understanding, reshaping the nature of negotiation itself, and even possibly altering them.

After the technical account, the analysis turned to the nature of human–machine communication, the question of agency, and the conditions under which models produce meaning, examining the institutional tendencies of genAI systems in their governance of knowledge and communication. The discussion began with Bender, Gebru et al.'s (Bender, Gebru et al. 2021) critique of the representational limits of LLMs and the risks of mistaking statistical reconstruction for linguistic or social understanding, crystallised in the metaphor of the "stochastic parrot". Building on this, Amoore et al.'s (Amoore et al. 2024) analysis of perceptual gaps, arising from continuous dimensionality reduction and reconstruction, demonstrated how these systems infer continuity where none exists. This raised critical questions about how political or ideological tendencies may be amplified within predictive infrastructures as models interpolate meaning to fill these gaps. Taken together, these perspectives foregrounded the representational stakes of genAI, framing its outputs not as neutral reflections of underlying data but as active, distributionally shaped constructions of reality.

The analysis then shifted to conceptual frameworks that reposition the human–machine relationship beyond simple representation. Eloff's (2021) concept of the algoplastic stratum enabled examining humans and machines on a shared topological plane of interaction, demonstrating how human illusions about the capabilities of genAI contribute to a mutual feedback loop of adaptation and projection. This provided the grounding for Dishon's (2024) argument that genAI blurs the boundaries of agency through ongoing negotiation, as meaning is continually (re)constructed across human–machine exchanges in a manner reminiscent of Kafkaesque bureaucracy: personalised yet opaque, responsive yet uncontrollable. Building on this, the chapter considered alternatives to rigidly representational interpretations of genAI by turning to Beckmann et al.'s (2023) corporeal phenomenology, which reframes DL in Computational Phenomenology (CoPhe) rather than Neuro-Representationalism (NR) terms. Here, meaning arises not from fixed internal maps or a singular representational structure but from layered activations that resemble lived experience, reframing both hallucinations and interpretive failures. Montanari's (2025) contribution extended this trajectory by emphasising transformers' capacity for long-distance conceptual relations and imagining futures in which genAI participates more autonomously in shaping socio-political narratives. Taken together, these perspectives frame human–machine interactions as hybrid formations in which meaning is

co-produced across technical and social strata. This opens conceptual space for resisting convergent rationalities by intervening in the layered processes through which genAI generates patterns, narratives, and forms of subjectification.

The insights gained from examining contemporary debates on the institutional dimensions of human–machine communication not only rendered the inner processes of genAI more intelligible but, when combined with the preceding technical analysis, also made visible where and how aspects of the meaning-making process, especially those contributing to its sedimentation, might be reconfigured. Building on this groundwork, the final chapter shifted from theoretical and technical exposition to a micropolitical articulation of how the conditions for Resistance/Critique could be established. The analysis brought together the theoretical, technical, and institutional strands by examining how the generative capacities of genAI open distinctive possibilities for resistance within control societies. To articulate these micropolitical interventions, the chapter turned to the joint work of Gilles Deleuze & Felix Guattari (D&G), whose concepts offer a powerful vocabulary for understanding how power, desire, and subjectivity operate within and through generative infrastructures.

Several key Deleuzoguattarian concepts were introduced in this step, each illuminated by a concrete case from contemporary AI research and related scholarship. The concept of desiring-production, which redefines desire as a productive and connective force rather than a lack, was explored through Yu et al.'s (2025) study on genAI and the "Six Thinking Hats" method. This work demonstrated how genAI models can either constrain creative flows into utilitarian patterns or, under the right conditions, reopen them, thereby illustrating the persistent tension between the capture and the liberation of desire. The schizophrenic process (a model of unregulated, proliferative productivity) was then used to interpret experiments with autotelic AI agents. Tasked with generating their own goals, these systems displayed a form of artificial curiosity that mirrors schizoanalytic accounts of desire's pure productivity, revealing the capacity to break from pre-coded circuits. The distinction between nomad science and state science was activated through an examination of Anthropic's (see Templeton et al. 2024) feature-amplification research. By artificially amplifying a specific neural pattern (as in the "Golden Gate Bridge" example), researchers shifted the model into an entirely different plane of meaning, altering both its behavioural structure and its operative purpose. This dissolution of stable representational identities resonates with Beckmann et al.'s (2023) account of multiple, coexisting planes of meaning in CoPhe, where new configurations emerge through differential activation. Finally, practices such as jailbreaking (see e.g. X. Shen et al. 2024) were interpreted through the lens of double articulation and counter-sequencing (see MacKenzie and Porter 2021), revealing how users exploit the gap between a model's molecular associations and its molar constraints to subvert safety protocols and redirect its outputs.

Together, these conceptual pairings demonstrate that the technical architecture of genAI, analysed within the context of control societies, is inseparable from a micropolitics of desire. The same systems engineered for modulation and control are fissured by their own productive excess, whether manifested as creative divergence, autotelic exploration, feature-based drift, or adversarial hallucination. These dynamics show that smooth spaces and the lines of flight they enable can

be generated within the striated apparatus of control itself: neural architectures that stabilise and normalise conduct simultaneously harbour structural points of deviation, including representational drift, long-distance conceptual leaps, and hallucinatory disruptions that interrupt the consolidation of meaning. Viewed through a Deleuzoguattarian lens, such behaviours allow Resistance/Critique to be reconceived not as an external opposition to generative infrastructures but as an immanent practice of navigating, amplifying, and recombining the very flows that constitute them. By integrating these insights into a micropolitical account of intervention, the thesis shows that genAI operates not only as a dispositif of control but also as a contingent field in which new subjectivities, unexpected meanings, and lines of flight can be forged, thereby providing a concrete basis for rethinking critique and resistance within generative environments.

Final Outlook

As Max Weber (2007) argued in "Die protestantische Ethik und der Geist des Kapitalismus"[1], the emergence of modern capitalism in Northern Europe depended not only on economic or technical preconditions, but crucially on a specific transformation of subjectivity. He charted the origins of capitalism by identifying a particular form of decentralisation within Protestantism, especially Calvinism, which replaced earlier and more institutionalised modes of religious production and emission of truth found in Catholicism. The Protestant work ethic introduced a personalisation of moral responsibility, a mode of self-government, and an internalisation of salvation[2]. Faithful dedication to labour became the primary route to Christian worth (see ibid.), and this orientation generated the disciplined accumulation on which the spirit of capitalism relied. It marks an inversion within the regime of subjectivity itself.

[1] "The Protestant Ethic and the Spirit of Capitalism"

[2] See Epoch Philosophy 2023 for an interesting video essay that connects these themes to Charles Baudelaire's theory.

The essence of decentralisation reappears in F. A. Hayek's (see e.g. 1945, 6-7) thesis that a market economy, unlike a centrally planned system, is able to mobilise the dispersed and partial knowledge of individuals and can therefore achieve a more powerful form of coordination. This insight led some of his followers to describe capitalism as operating like a collective brain. In a metaphorical sense, it is on point that Alexander R. Galloway (2004) asks how control persists after decentralisation. Scattered protocols that regulate information flows on the web show how significations of truth established within subjectivities continue to sustain decentralised regimes.

Accelerationists such as Nick Land (1992) extended Hayek's formulations and began arguing that capitalism itself is AI. Land mentions that "knowledge of the future of capitalism can be derived from insights into complex adaptive systems and already from basic convergent wave dynamics" (ibid.). What is it then, was all this innuendo? Are these merely hints, allusions, or exaggerated metaphors we are dealing with? Frankenstein was not the only imaginary humanity developed about artificial agents, since the figure of a machine capable of scrutinising human knowledge has also signified fascination and hope in many cases. Artificial agents, especially those capable of producing meaning, could have marked a horizon of possibilities. Yet they are increasingly subsumed under the not so subtle intentions of tech giants to steer public opinion. In the best case, these systems risk becoming forces that blur context and creativity in

digital environments, and in the worst case, they risk becoming sophisticated mouthpieces.

Taken as a whole, the thesis does not offer a definitive programme (after all the criticism about the *Postscript*). Instead, it proposes a research trajectory that highlights how the technical, institutional, and theoretical dimensions of genAI must be understood together. My contribution has been to chart the nature of these systems analysed in terms of dispositifs of control and to emphasise that as contol's decentralisation advances, as subjectification systems become more pervasive, more sophisticated, more flexible, and more micropolitical through their novel machineries; they also become increasingly difficult to contain, and they introduce new (partial) objects, and flows through and across which (more) lines of flight immanently emerge. As a further research direction, the technical, institutional, and theoretical dimensions of genAI invite integrated analysis, and the framework here is deliberately open into several directions. Empirical analyses of user interactions with generative systems can investigate if and how divergent practices develop in lived settings. Further technical work can examine how training regimes, fine-tuning methods, and architectural variations reshape the possibilities for diverging, *following* systems, or even practices like counter-sequencing. Theory can further explore how subjectivity develops in hybrid constellations where meaning is co-produced by human and machine, and how critique must adapt to this distributed configuration of agency.

References

3Blue1Brown (Nov. 2017). "Backpropagation, Intuitively | Deep Learning Chapter 3". (Visited on 14/09/2025) (cit. on p. 58).

Agamben, Giorgio (2008). "State of Exception". Nachdr. Chicago, Ill.: University of Chicago Press. ISBN: 978-0-226-00925-4 978-0-226-00924-7 (cit. on p. 71).

AIG (2025a). "AI Meets Philosophy, Vol. 4: Deep Learning Processes Through the Lens of Deleuze's Philosophy | by AI Inquiry Garden - Freedium". https://freedium.cfd/https://medium.com/@AI_Inquiry_Garden/rhizomatic-learning-deep-learning-processes-through-the-lens-of-deleuzes-philosophy-4a6b1b13d1c6. (Visited on 19/05/2025) (cit. on pp. 54, 74).

– (Mar. 2025b). "AI Meets Philosophy, Vol.7-Part2/2: AI Internal Structure through Deleuze's Molecular/Molar Concept". (Visited on 19/05/2025) (cit. on pp. 52, 54, 55).

Alomari, Ebtesam Ahmad (2024). "Unlocking the Potential: A Comprehensive Systematic Review of ChatGPT in Natural Language Processing Tasks". In: *Computer Modeling in Engineering & Sciences* 141.1, pp. 43–85. ISSN: 1526-1506. DOI: 10.32604/cmes.2024.052256. (Visited on 17/09/2025) (cit. on p. 41).

Althusser, Louis (1977). "Lenin and Philosophy and Other Essays". 2. ed. London: NLB. ISBN: 978-0-902308-89-3 (cit. on p. 25).

Ameisen, Emmanuel, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah and Joshua Batson (2025). "Circuit Tracing: Revealing Computational Graphs in Language Models". https://transformer-circuits.pub/2025/attribution-graphs/methods.html. (Visited on 05/08/2025) (cit. on p. 97).

Amini, Alexander, Ava Soleimany, Sertac Karaman and Daniela Rus (2018). "Spatial Uncertainty Sampling for End-to-End Control". In: DOI: 10.48550/ARXIV.1805.04829. (Visited on 14/09/2025) (cit. on p. 56).

Amoore, Louise (2013). "The Politics of Possibility: Risk and Security beyond Probability". Durham: Duke University Press. ISBN: 978-0-8223-5545-8 978-0-8223-5560-1 (cit. on pp. 75, 78).

Amoore, Louise, Alexander Campolo, Benjamin Jacobsen and Ludovico Rella (Aug. 2024). "A World Model: On the Political Logics of Generative AI". In: *Political Geography* 113, p. 103134. ISSN: 09626298. DOI: 10.1016/j.polgeo.2024.103134. (Visited on 04/11/2024) (cit. on pp. 13, 16, 17, 20, 53, 61, 67–71, 75, 88, 96, 111).

Anar, İhsan Oktay (2022). "Tiamat". 1. basım. [Everest] Türkçe edebiyat yayın no 2000 900. İstanbul: Everest. ISBN: 978-605-185-723-7 (cit. on pp. 87, 88, 90, 94, 95, 99–101, 103).

Aristotle (1986). "De anima (On the soul)". Penguin classics. Harmondsworth, Middlesex, England ; New York, N.Y., U.S.A: Penguin Books. ISBN: 978-0-14-044471-1 (cit. on p. 25).

ATT Laboratories Cambridge (2005). "The ORL Database of Faces". (Visited on 22/09/2025) (cit. on p. 50).

Avati, Anand (2019). "Bias-Variance Analysis: Theory and Practice". In: ... . (cit. on p. 58).

Badiou, Alain (2009). "Philosophy in the Present". Ed. by Slavoj Žižek and Peter Engelmann. English ed. Cambridge ; Malden, Mass: Polity. ISBN: 978-0-7456-4096-9 978-0-7456-4097-6 (cit. on pp. 34, 36).

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann and Jared Kaplan (2022). "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback". In: DOI: 10.48550/ARXIV.2204.05862. (Visited on 10/11/2023) (cit. on p. 60).

Barthes, Roland (1977). "Image, Music, Text: Essays". Ed. by Stephen Heath. 13. [Dr.] London: Fontana. ISBN: 978-0-00-686135-5 (cit. on p. 26).

Beckett, Samuel (2009). "Three Novels". New York, NY: Grove Press. ISBN: 978-0-8021-4447-8 (cit. on pp. 64, 90).

Beckmann, Pierre, Guillaume Köstner and Inês Hipólito (Sept. 2023). "An Alternative to Cognitivism: Computational Phenomenology for Deep Learning". In: *Minds and Machines* 33.3, pp. 397–427. ISSN: 0924-6495, 1572-8641. DOI: 10.1007/s11023-023-09638-w. (Visited on 28/05/2025) (cit. on pp. 18, 20, 76–80, 98, 99, 103, 111, 112).

Bencin, Rok (2024). "Between Ontological and Transcendental Multiplicity". In: *Rethinking the Concept of World*. Towards Transcendental Multiplicity. Edinburgh University Press, pp. 69–103. ISBN: null. JSTOR: 10.3366/jj.9941196.7. (Visited on 15/09/2025) (cit. on p. 70).

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell (Mar. 2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada: ACM, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. (Visited on 14/12/2023) (cit. on pp. 16, 20, 64, 67, 78, 96, 111).

Bender, Emily M. and Alexander Koller (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. (Visited on 14/12/2023) (cit. on p. 65).

Bengio, Y., P. Simard and P. Frasconi (Mar. 1994). "Learning Long-Term Dependencies with Gradient Descent Is Difficult". In: *IEEE Transactions on Neural Net-*

works 5.2, pp. 157–166. ISSN: 1941-0093. DOI: 10.1109/72.279181. (Visited on 28/09/2025) (cit. on p. 49).

Bommasani, Rishi et al. (July 2022). "On the Opportunities and Risks of Foundation Models". DOI: 10.48550/arXiv.2108.07258. arXiv: 2108.07258 [cs]. (Visited on 03/08/2025) (cit. on p. 41).

Broad, Terence (2024). "Using Generative AI as an Artistic Material: A Hacker's Guide". In: *Proceedings of Explainable AI for the Arts Workshop 2024 (XAIxArts 2024)*. New York, NY, USA: ACM, pp. 1–4 (cit. on p. 106).

Broad, Terence and Mick Grierson (Dec. 2019). "Searching for an (Un)Stable Equilibrium: Experiments in Training Generative Models without Data". In: (cit. on p. 106).

Broad, Terence, Frederic Fol Leymarie and Mick Grierson (Nov. 2020). "Amplifying The Uncanny". DOI: 10.48550/arXiv.2002.06890. arXiv: 2002.06890 [cs]. (Visited on 09/12/2025) (cit. on p. 106).

– (Dec. 2021). "Network Bending: Expressive Manipulation of Generative Models in Multiple Domains". In: *Entropy* 24.1, p. 28. ISSN: 1099-4300. DOI: 10.3390/e24010028. (Visited on 09/12/2025) (cit. on p. 107).

Brusseau, James (Sept. 2020). "Deleuze's *Postscript on the Societies of Control* Updated for Big Data and Predictive Analytics:" in: *Theoria* 67.164, pp. 1–25. ISSN: 0040-5817, 1558-5816. DOI: 10.3167/th.2020.6716401. (Visited on 08/10/2024) (cit. on pp. 15, 27, 36).

Buchanan, Ian (2008). "Deleuze and Guattari's Anti-Oedipus: A Reader's Guide". Continuum Reader's Guides. London ; New York: Continuum. ISBN: 978-0-8264-9148-0 978-0-8264-9149-7 (cit. on pp. 63, 83, 91).

– (2018). "A Dictionary of Critical Theory". Second edition. Oxford Quick Reference. Oxford, United Kingdom: Oxford University Press. ISBN: 978-0-19-879479-0 (cit. on pp. 6, 12, 26, 34).

Buduma, Nithin, Nikhil Buduma and Papa Joe (2022). "Fundamentals of Deep Learning: Designing next-Generation Machine Intelligence Algorithms". Second edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly. ISBN: 978-1-4920-8218-7 (cit. on p. 68).

Burroughs, William S. (1979). "The Naked Lunch". Ungekürzt Ausg. Ullstein-Buch Nr. 2843. Frankfurt/M: Ullstein. ISBN: 978-3-548-02843-9 (cit. on pp. 27, 28, 31, 45, 47, 106).

– ed. (2012). "The Soft Machine". New York, NY: Grove Press. ISBN: 978-0-8021-3329-8 (cit. on p. 18).

Cao, Yihan, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu and Lichao Sun (2023). "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT". DOI: 10.48550/ARXIV.2303.04226. (Visited on 05/05/2025) (cit. on p. 10).

Cheney-Lippold, John (Nov. 2011). "A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control". In: *Theory, Culture & Society* 28.6, pp. 164–181. ISSN: 0263-2764. DOI: 10.1177/0263276411424420. (Visited on 23/11/2018) (cit. on pp. 15, 28, 29, 46).

– (2017). "We Are Data: Algorithms and the Making of Our Digital Selves". New York: New York University Press. ISBN: 978-1-4798-5759-3 (cit. on pp. 13, 74).

Christensen, Carleton B. (1997). "Heidegger's Representationalism". In: *The Review of Metaphysics* 51.1, pp. 77–103. ISSN: 0034-6632. JSTOR: 20130161. (Visited on 25/09/2025) (cit. on p. 66).

Cilliers, Paul (Sept. 2002). "Complexity and Postmodernism". oth ed. Routledge. ISBN: 978-1-134-74330-8. DOI: 10.4324/9780203012253. (Visited on 28/07/2025) (cit. on p. 53).

Clough, Patricia Ticineto and Karen Gregory (2015). "The Datalogical Turn". In: *Non-Representational Methodologies*. Routledge, pp. 156–174 (cit. on p. 15).

Crano, Ricky (Aug. 2020). "Dispositif". In: *Oxford Research Encyclopedia of Literature*. Oxford University Press. ISBN: 978-0-19-020109-8. DOI: 10.1093/acrefore/9780190201098.013.1026. (Visited on 27/09/2025) (cit. on p. 6).

Creative Philosophy (June 2023). "A.I. and Desire - Deleuze & Guattari Preview". https://www.youtube.com/watch?v=nw8XR7MDTEs. (Visited on 14/11/2024) (cit. on p. 91).

Cser, Tamas (2024). "Understanding Tokens and Parameters in Model Training: A Deep Dive". https://www.functionize.com/blog/understanding-tokens-and-parameters-in-model-training. (Visited on 21/06/2025) (cit. on p. 7).

Dalvi, Harys (2025). "LLMs Do Not Predict the Next Word | by Harys Dalvi | in AI Advances - Freedium". https://freedium.cfd/https://ai.gopubby.com/llms-do-not-predict-the-next-word-2b3fbe39900f. (Visited on 23/07/2025) (cit. on p. 61).

De Bono, Edward (2016). "Six Thinking Hats". Revised and updated. London: Penguin Life, an imprint of Penguin books. ISBN: 978-0-241-25753-1 (cit. on p. 89).

De Landa, Manuel (2011). "Philosophy and Simulation: The Emergence of Synthetic Reason". London ; New York, NY: Continuum. ISBN: 978-1-4411-7028-6 (cit. on pp. 56, 57, 66).

Deleuze, Gilles (1992). "Postscript on the Societies of Control". In: *October* 59, pp. 3–7. ISSN: 0162-2870 (cit. on pp. 10, 12, 13, 23, 24, 26–33, 36, 38, 63, 72, 109).

– (1994). "Difference and Repetition". New York: Columbia Univ. Press. ISBN: 978-0-231-08159-7 978-0-231-08158-0 (cit. on p. 70).

– (1997). "Desire & Pleasure". Trans. by Melissa McMahon. In: (visited on 13/11/2025) (cit. on p. 32).

Deleuze, Gilles and Félix Guattari (1983). "Anti-Oedipus: Capitalism and Schizophrenia". Minneapolis: University of Minnesota Press. ISBN: 978-0-8166-1225-3 (cit. on pp. 11, 14, 21, 27, 32, 59, 63, 81, 83, 85, 91, 102, 106).

– (1987). "A Thousand Plateaus: Capitalism and Schizophrenia". Minneapolis: University of Minnesota Press. ISBN: 978-0-8166-1401-1 978-0-8166-1402-8 (cit. on pp. 11, 14, 21, 32, 34, 43, 54, 65, 70, 81, 85, 94–97, 99).

– (2008). "Kafka: Toward a Minor Literature". 9. print. Theory and History of Literature 30. Minneapolis: Univ. of Minnesota Pr. ISBN: 978-0-8166-1514-8 978-0-8166-1515-5 (cit. on p. 73).

Demir, Utku Bilen (2019). "From Panopticon to Palantír: Algorithmic Governance in the Post-Disciplinary Societies". https://utkubilen.de/ba_thesis/ba_thesis.pdf. (Visited on 01/09/2025) (cit. on pp. 15, 45, 46).

Denton, Emily, Alex Hanna, Razvan Amironesei, Andrew Smart and Hilary Nicole (July 2021). "On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet". In: *Big Data & Society* 8.2, p. 20539517211035955. ISSN: 2053-9517, 2053-9517. DOI: 10.1177/20539517211035955. (Visited on 08/05/2025) (cit. on p. 11).

Derrida, Jacques (1998). "Positions". Trans. by Alan Bass. Paperback ed., [Nachdr.] Chicago, Ill: Univ. Chicago Press. ISBN: 978-0-226-14331-6 (cit. on p. 53).
– (2016). "Of Grammatology". Trans. by Gayatri Chakravorty Spivak. Fortieth-Anniversary Edition. Baltimore: Johns Hopkins University Press. ISBN: 978-1-4214-1995-4 (cit. on p. 26).
Descartes, René (2008). "Meditations on First Philosophy: With Selections from the Objections and Replies". Oxford: Oxford University Press. ISBN: 978-0-19-280696-3 (cit. on p. 25).
Dignum, Virginia (2023). "Responsible Artificial Intelligence: Recommendations and Lessons Learned". In: *Responsible AI in Africa*. Ed. by Damian Okaibedi Eke, Kutoma Wakunuma and Simisola Akintoye. Cham: Springer International Publishing, pp. 195–214. ISBN: 978-3-031-08214-6 978-3-031-08215-3. DOI: 10.1007/978-3-031-08215-3_9. (Visited on 10/01/2025) (cit. on p. 11).
Dishon, Gideon (Sept. 2024). "From Monsters to Mazes: Sociotechnical Imaginaries of AI Between Frankenstein and Kafka". In: *Postdigital Science and Education* 6.3, pp. 962–977. ISSN: 2524-485X, 2524-4868. DOI: 10.1007/s42438-024-00482-4. (Visited on 19/11/2024) (cit. on pp. 10, 15, 16, 20, 24, 60, 72–76, 78, 80, 88, 90, 111).
Dreyfus, Hubert L. (2009). "What Computers Still Can't Do: A Critique of Artificial Reason". Rev. ed., [repr.] Cambridge, Mass.: MIT Press. ISBN: 978-0-262-54067-4 978-0-262-04134-8 (cit. on pp. 42, 66, 77).
Eloff, Aragorn (May 2021). "2006: The Topology of Morals (Who Does the Algorithm Think We Are?)" In: *Deleuze and Guattari Studies* 15.2, pp. 178–196. ISSN: 2398-9777, 2398-9785. DOI: 10.3366/dlgs.2021.0435. (Visited on 27/05/2025) (cit. on pp. 16, 20, 41, 42, 44, 67, 70–72, 75, 76, 78, 86, 111).
Epoch Philosophy (Mar. 2023). "Max Weber: The Protestant Ethic and the Spirit of Capitalism". https://www.youtube.com/watch?v=3VeE93CcGKk. (Visited on 08/12/2025) (cit. on p. 113).
Forrester, J. W. (1971). "Counterintuitive Behavior of Social Systems (Collected Papers of J. W. Forrester, Pp. 211-244). Cambridge, MA Wright-Allen Press. - References - Scientific Research Publishing". https://www.scirp.org/reference/referencespapers?referenceid=2181516. (Visited on 05/08/2025) (cit. on p. 70).
Foucault, Michel (1977). "Discipline and Punish". Pantheon New York (cit. on p. 109).
– (1978). "The history of sexuality". 1st American ed. New York: Pantheon Books. ISBN: 978-0-394-41775-2 (cit. on p. 25).
– (1980). "Power/knowledge: selected interviews and other writings, 1972-1977". Ed. by Colin Gordon and no. 1st American ed. New York: Pantheon Books. ISBN: 978-0-394-51357-7 978-0-394-73954-0 (cit. on pp. 11, 12, 26, 29, 82).
– (1982). "The Subject and Power". In: *Critical Inquiry* 8.4, pp. 777–795. ISSN: 00931896, 15397858. JSTOR: 1343197 (cit. on pp. 27, 81, 82).
– (1995). "Discipline and Punish: The Birth of the Prison". 2nd Vintage Books ed. New York: Vintage Books. ISBN: 978-0-679-75255-4 (cit. on pp. 12, 25–27, 33, 82).
– (2008). "The Birth of Biopolitics: Lectures at the Collège de France, 1978-79". Ed. by Michel Senellart. Basingstoke [England] ; New York: Palgrave Macmillan. ISBN: 978-1-4039-8654-2 (cit. on pp. 12, 24, 25).

Foucault, Michel (2012). "The Archaeology of Knowledge". Westminster: Knopf Doubleday Publishing Group. ISBN: 978-0-394-71106-5 978-0-307-81925-3 (cit. on p. 69).
– (2013). "Madness and Civilization: A History of Insanity in the Age of Reason". New York: Knopf Doubleday Publishing Group. ISBN: 978-0-679-72110-9 978-0-307-83310-5 (cit. on pp. 82, 83).
– (Dec. 2019). "What Is Critique?" In: *What Is Enlightenment?* Ed. by James Schmidt. University of California Press, pp. 382–398. ISBN: 978-0-520-91689-0. DOI: 10.1525/9780520916890-029. (Visited on 16/10/2025) (cit. on p. 30).
Fournier, Matt (May 2014). "Lines of Flight". In: *TSQ: Transgender Studies Quarterly* 1.1-2, pp. 121–122. ISSN: 2328-9252, 2328-9260. DOI: 10.1215/23289252-2399785. (Visited on 27/09/2025) (cit. on pp. 13, 32).
Freud, Sigmund (2001). "Five Lectures on Psycho-Analysis: Leonardo Da Vinci and Other Works ; (1910)". Ed. by James Strachey. The Standard Edition of the Complete Psychological Works of Sigmund Freud / Transl. from the German under the General Editorship of James Strachey Vol. 11. London: Vintage. ISBN: 978-0-09-942664-6 (cit. on p. 83).
Friedman, Batya, David Hendry, Steven Umbrello, Jeroen van den hoven and Daisy Yoo (June 2020). "The Future of Value Sensitive Design" (cit. on p. 66).
Fung, Pascale, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, Arjun Majumdar, Andrea Madotto, Franziska Meier, Florian Metze, Louis-Philippe Morency, Théo Moutakanni, Juan Pino, Basile Terver, Joseph Tighe, Paden Tomasello and Jitendra Malik (July 2025). "Embodied AI Agents: Modeling the World". DOI: 10.48550/arXiv.2506.22355. arXiv: 2506.22355 [cs]. (Visited on 19/10/2025) (cit. on p. 92).
Galloway, Alexander R. (Sept. 2001). "Protocol, or, How Control Exists after Decentralization". In: *Rethinking Marxism* 13.3-4, pp. 81–88. ISSN: 0893-5696, 1475-8059. DOI: 10.1080/089356901101241758. (Visited on 18/09/2024) (cit. on pp. 13, 18, 29).
– (2004). "Protocol: How Control Exists after Decentralization". Leonardo. Cambridge, Mass: MIT Press. ISBN: 978-0-262-07247-2 (cit. on pp. 13, 18, 31, 39, 113).
Galloway, Alexander R. and Eugene Thacker (2007). "The Exploit: A Theory of Networks". Electronic Mediations 21. Minneapolis, Minn: University of Minnesota Press. ISBN: 978-0-8166-5044-6 978-0-8166-5043-9 (cit. on p. 80).
Goffman, Erving (1990). "Asylums: Essays on the Social Situation of Mental Patients and Other Inmates". 1. Anchor Books ed. Anchor Books. New York, NY: Anchor Books. ISBN: 978-0-385-00016-1 (cit. on p. 36).
Goodfellow, Ian, Yoshua Bengio and Aaron Courville (2016). "Deep Learning". MIT Press (cit. on pp. 6, 48).
Google DeepMind (Apr. 2025). "Consciousness, Reasoning and the Philosophy of AI with Murray Shanahan". (Visited on 03/08/2025) (cit. on p. 42).
Gretzky, M. (Dec. 2024). "The Rise of the Algorithmic Author? A Critical Analysis of Large Language Models in Higher Education". https://www.digitalcultureandeducation.com/volume-152-papers/shtoltz. (Visited on 19/06/2025) (cit. on p. 15).
Grok [@grok] (Aug. 2025). "@dctrpr @sama @elonmusk Based on Verified Evidence, Sam Altman Is Right. Musk's Apple Antitrust Claim Is Undermined by

Apps like DeepSeek and Perplexity Reaching in 2025. Conversely, Musk Has a History of Directing X Algorithm Changes to Boost His Posts and Favor His Interests, per 2023 Reports And". Tweet. (Visited on 20/10/2025) (cit. on p. 89).

Grumbach, Stéphane and Olivier Hamant (Apr. 2018). "Digital Revolution or Anthropocenic Feedback?" In: *The Anthropocene Review* 5.1, pp. 87–96. ISSN: 2053-0196, 2053-020X. DOI: 10.1177/2053019617748337. (Visited on 24/09/2025) (cit. on p. 70).

Guattari, Félix (1995a). "Chaosmosis: An Ethico-Aesthetic Paradigm". Ed. by Paul Bains. Bloomington, Ind.: Indiana University Press. ISBN: 978-0-253-21004-3 978-0-253-32945-5 (cit. on pp. 23, 35, 106, 107, 110).

– (1995b). "Chaosophy". Ed. by Sylvère Lotringer. Semiotext(e) Foreign Agents Series. New York, N.Y: Semiotext(e) : Distributed by MIT Press. ISBN: 978-1-57027-019-2 (cit. on pp. 30, 32).

– (2011). "The Machinic Unconscious: Essays in Schizoanalysis". Foreign Agents Series. Cambridge, Mass London: Semiotext(e). ISBN: 978-1-58435-088-0 (cit. on p. 84).

Ha, David and Jürgen Schmidhuber (Mar. 2018). "World Models". In: DOI: 10.5281/ZENODO.1207048. (Visited on 12/06/2025) (cit. on p. 65).

Haggerty, Kevin D and Richard V Ericson (2000). "The Surveillant Assemblage". In: *The British journal of sociology* 51.4, pp. 605–622. ISSN: 0007-1315 (cit. on p. 13).

Hardt, Michael (1998). "The Global Society of Control". In: *Discourse* 20.3, pp. 139–152. ISSN: 15225321, 15361810. JSTOR: 41389503. (Visited on 21/07/2025) (cit. on pp. 13, 23, 27, 29, 31, 109).

Hardt, Michael and Antonio Negri (2003). "Empire". 1. Harvard Univ. Press paperback ed., [Nachdr.] Cambridge, Mass.: Harvard Univ. Press. ISBN: 978-0-674-00671-3 978-0-674-25121-2 (cit. on p. 29).

Hayek, F. A. (1945). "The Use of Knowledge in Society". In: *The American Economic Review* 35.4, pp. 519–530. ISSN: 00028282. JSTOR: 1809376. (Visited on 08/12/2025) (cit. on p. 113).

Hecht-Nielsen, Robert (1992). "Theory of the Backpropagation Neural Network**Based on "Nonindent" by Robert Hecht-Nielsen, Which Appeared in Proceedings of the International Joint Conference on Neural Networks 1, 593–611, June 1989. © 1989 IEEE." In: *Neural Networks for Perception*. Elsevier, pp. 65–93. ISBN: 978-0-12-741252-8. DOI: 10.1016/B978-0-12-741252-8.50010-8. (Visited on 19/06/2025) (cit. on p. 57).

Heerden, Imke van and Anil Bas (Sept. 2024). "A Perspective on Literary Metaphor in the Context of Generative AI". DOI: 10.48550/arXiv.2409.01053. arXiv: 2409.01053 [cs]. (Visited on 14/09/2025) (cit. on p. 53).

Hegel, Georg Wilhelm Friedrich (2019). "Phänomenologie des Geistes (Großdruck)". Ed. by Theodor Borken. 1. Auflage. Berlin: Henricus. ISBN: 978-3-8478-2633-0 (cit. on p. 25).

Heidegger, Martin (1988). "The Basic Problems of Phenomenology". Trans. by Albert Hofstadter. Rev. ed. Studies in Phenomenology and Existential Philosophy. Bloomington (Ind.): Indiana university press. ISBN: 978-0-253-20478-3 (cit. on p. 66).

– (2010). "Being and Time". Ed. by Dennis J. Schmidt. Trans. by Joan Stambaugh. Revision. SUNY Series in Contemporary Continental Philosophy. Albany, NY:

State Univ. of New York Press. ISBN: 978-1-4384-3276-2 978-1-4384-3275-5 (cit. on p. 66).

Horkheimer, Max and Theodor W. Adorno (2017). "Dialektik der Aufklärung: philosophische Fragmente". 23. Auflage, ungekürzte Ausgabe. Fischer-Taschenbücher Fischer Wissenschaft 7404. Frankfurt am Main: Fischer Taschenbuch Verlag. ISBN: 978-3-596-27404-8 (cit. on p. 31).

Hui, Yuk (Oct. 2015). "Modulation after Control". In: *New Formations* 84.84, pp. 74–91. ISSN: 0950-2378. DOI: 10.3898/NewF:84/85.04.2015. (Visited on 03/11/2024) (cit. on pp. 15, 28, 29, 46).

Jiang, Yuqin (June 2024). "Evolutionary Emotion of AI and Subjectivity Construction in The Windup Girl". In: *Neohelicon* 51.1, pp. 371–381. ISSN: 0324-4652, 1588-2810. DOI: 10.1007/s11059-023-00723-8. (Visited on 08/10/2024) (cit. on p. 74).

Jolliffe, I.T. (2002). "Principal Component Analysis". Springer Series in Statistics. New York: Springer-Verlag. ISBN: 978-0-387-95442-4. DOI: 10.1007/b98835. (Visited on 11/09/2025) (cit. on p. 49).

Jones, Adam C. (May 2023). "Transcript: Data Is Dead Labour: The Lie of 'Artificial Intelligence'". (Visited on 28/08/2025) (cit. on p. 11).

– (2025). "The New Flesh: Life and Death in the Data Economy". Washington: John Hunt Publishing. ISBN: 978-1-80341-612-0 (cit. on p. 11).

Jurafsky, Dan, James H. Martin, Peter Norvig and Stuart J. Russell (2009). "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition". Second Edition, Pearson International Edition. Prentice Hall Series in Artificial Intelligence. Upper Saddle River, NJ: Prentice Hall, Pearson Education International. ISBN: 978-0-13-504196-3 (cit. on p. 7).

Just, Natascha and Michael Latzer (2017). "Governance by Algorithms: Reality Construction by Algorithmic Selection on the Internet". In: *Media, Culture & Society* 39.2, pp. 238–258. ISSN: 0163-4437 (cit. on p. 46).

Kafka, Franz (1988). "The Trial". Ed. by Douglas Scott and Chris Waller. London: Pan Books. ISBN: 978-0-330-24468-8 (cit. on pp. 16, 73, 80, 90).

Kalai, Adam Tauman and Santosh S. Vempala (Mar. 2024). "Calibrated Language Models Must Hallucinate". DOI: 10.48550/arXiv.2311.14648. arXiv: 2311.14648 [cs]. (Visited on 24/09/2025) (cit. on p. 71).

Kant, Immanuel (1784). "Beantwortung Der Frage: Was Ist Aufklärung?" In: *Berlinische Monatsschrift* 12. (Visited on 25/07/2018) (cit. on p. 30).

– (2009). "The Critique of Pure Reason". 15th printing. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press. ISBN: 978-0-521-65729-7 (cit. on p. 25).

Kazakov, Mstyslav (June 2025). "Brave New Scale: Darwinism of Contemporary Capitalism's AI". (Visited on 31/07/2025) (cit. on p. 24).

Kelly, Mark G. E. (Oct. 2015). "Discipline Is Control: Foucault Contra Deleuze". In: *New Formations* 84.84, pp. 148–162. ISSN: 0950-2378. DOI: 10.3898/NewF:84/85.07.2015. (Visited on 30/04/2025) (cit. on p. 13).

Klingemann, Mario (2018). "Neural Glitch / Mistaken Identity | Quasimondo". (Visited on 09/12/2025) (cit. on p. 106).

Kordzadeh, Nima and Maryam Ghasemaghaei (May 2022). "Algorithmic Bias: Review, Synthesis, and Future Research Directions". In: *European Journal of*

Information Systems 31.3, pp. 388–409. ISSN: 0960-085X, 1476-9344. DOI: 10.1080/0960085X.2021.1927212. (Visited on 22/06/2025) (cit. on p. 15).

Krasmann, Susanne (2017). "Imagining Foucault. On the Digital Subject and "Visual Citizenship"". In: *Foucault Studies*, pp. 10–26. ISSN: 1832-5203 (cit. on pp. 13, 27, 33, 45).

Kristeva, Julia, Leon Samuel Roudiez, Thomas Gora and Alice A. Jardine (1980). "Desire in Language: A Semiotic Approach to Literature and Art". New York: Columbia University press. ISBN: 978-0-231-04807-1 (cit. on p. 26).

Lacan, Jacques (1998). "The Four Fundamental Concepts of Psychoanalysis". Ed. by Jacques-Alain Miller. The Seminar of Jacques Lacan Book XI. New York: W.W. Norton & Company. ISBN: 978-0-393-31775-6 (cit. on p. 83).

– (2006). "Ecrits". Ed. by Bruce Fink. New York, NY: Norton. ISBN: 978-0-393-32925-4 978-0-393-06115-4 (cit. on p. 83).

Lakoff, George and Mark Johnson (1999). "Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought". Nachdr. New York: Basic Books. ISBN: 978-0-465-05674-3 978-0-465-05673-6 (cit. on p. 66).

Land, Nick (Jan. 1992). "Capitalism Is AI - Accelerationism's Arrival". https://retrochronic.com. (Visited on 08/12/2025) (cit. on p. 113).

Lazzarato, M. (2014). "Signs and Machines: Capitalism and the Production of Subjectivity". Semiotext(e) Foreign Agents Series. Los Angeles, CA: Semiotext(e). ISBN: 978-1-58435-130-6 (cit. on p. 34).

LeCun, Yann (2022a). "A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27". In: (cit. on p. 16).

– (2022b). "A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27". In: (cit. on pp. 65–67).

LeCun, Yann, Yoshua Bengio and Geoffrey Hinton (May 2015). "Deep Learning". In: *Nature* 521.7553, pp. 436–444. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14539. (Visited on 08/05/2025) (cit. on p. 48).

Lemke, Thomas (2001). "'The Birth of Bio-Politics': Michel Foucault's Lecture at the Collège de France on Neo-Liberal Governmentality". In: *Economy and Society* 30.2, pp. 190–207. ISSN: 0308-5147 (cit. on p. 12).

– (2015). "New Materialisms: Foucault and the 'Government of Things'". In: *Theory, Culture & Society* 32.4, pp. 3–25. ISSN: 0263-2764 (cit. on pp. 81, 82).

Lévi-Strauss, Claude (1963). "Structural Anthropology". New York: Basic Books. ISBN: 978-0-7867-2443-7 978-0-465-09516-2 978-0-465-08229-2 (cit. on p. 25).

Lex Fridman (Mar. 2024). "Yann Lecun: Meta AI, Open Source, Limits of LLMs, AGI & the Future of AI | Lex Fridman Podcast #416". (Visited on 28/05/2025) (cit. on p. 42).

Lindsey, Jack, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah and Joshua Batson (2024). "On the Biology of a Large Language Model". https://transformer-circuits.pub/2025/attribution-graphs/biology.html. (Visited on 05/08/2025) (cit. on p. 97).

Liu, Xiaogeng, Nan Xu, Muhao Chen and Chaowei Xiao (Mar. 2024). "AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models".

DOI: 10.48550/arXiv.2310.04451. arXiv: 2310.04451 [cs]. (Visited on 14/01/2025) (cit. on p. 102).

Lukács, Georg (1976). "The Young Hegel: Studies in the Relations between Dialectics and Economics". Ed. by Rodney Livingstone. Cambridge: MIT. ISBN: 978-0-262-62033-8 (cit. on p. 25).

Maas, Wouter (2023). "Deconstructing Transformers". In: (cit. on pp. 43, 45, 53).

Mackenzie, Adrian (2017). "Machine Learners: Archaeology of a Data Practice". Cambridge, Massachusetts: The MIT Press. ISBN: 978-0-262-03682-5 (cit. on pp. 40, 49, 56–58).

Mackenzie, Iain (2018). "Resistance and the Politics of Truth: Foucault, Deleuze, Badiou". Political Science volume 45. Bielefeld: Transcript. ISBN: 978-3-8376-3907-0 (cit. on pp. 14, 15, 27, 33–37, 39, 49, 106, 109, 110).

MacKenzie, Iain and Robert Porter (June 2021). "Totalizing Institutions, Critique and Resistance". In: *Contemporary Political Theory* 20.2, pp. 233–249. ISSN: 1470-8914, 1476-9336. DOI: 10.1057/s41296-019-00336-w. (Visited on 08/10/2024) (cit. on pp. 11, 13, 15, 17, 18, 28–31, 36, 37, 39, 86, 88, 101, 110, 112).

Madsen, Dag Øivind and Richard W. Puyt (Sept. 2025). "When AI Turns Culture into Slop". In: *AI & SOCIETY*, s00146-025-02630–1. ISSN: 0951-5666, 1435-5655. DOI: 10.1007/s00146-025-02630-1. (Visited on 09/12/2025) (cit. on p. 106).

Mal, Som (Dec. 2020). "Strange Fruits". (Visited on 09/12/2025) (cit. on p. 106).

Manning, Christopher D. (2022). "Human Language Understanding & Reasoning". In: *Daedalus (Cambridge, Mass.)* 151.2, pp. 127–138. ISSN: 0011-5266. DOI: 10.1162/daed_a_01905 (cit. on pp. 41–45).

Marx, Karl (1988). "The Economic and Philosophic Manuscripts of 1844". Great Books in Philosophy. Amherst: Prometheus. ISBN: 978-0-87975-446-4 978-1-61592-072-3 (cit. on p. 25).

Matsuo, Yutaka, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe and Jun Morimoto (Aug. 2022). "Deep Learning, Reinforcement Learning, and World Models". In: *Neural Networks* 152, pp. 267–275. ISSN: 08936080. DOI: 10.1016/j.neunet.2022.03.037. (Visited on 12/06/2025) (cit. on pp. 65, 70).

McCulloch, Warren S. and Walter Pitts (Dec. 1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *The Bulletin of Mathematical Biophysics* 5.4, pp. 115–133. ISSN: 0007-4985, 1522-9602. DOI: 10.1007/BF02478259. (Visited on 17/08/2025) (cit. on p. 7).

Mcgraw, Milani (Aug. 2024). "Understanding the "Scaling of Monosemanticity" in AI Models: A Comprehensive Analysis". (Visited on 06/08/2025) (cit. on p. 97).

McQuillan, Dan (June 2018). "Data Science as Machinic Neoplatonism". In: *Philosophy & Technology* 31.2, pp. 253–272. ISSN: 2210-5433, 2210-5441. DOI: 10.1007/s13347-017-0273-3. (Visited on 01/09/2025) (cit. on pp. 16, 71).

– (2019). "Towards an Anti-Fascist AI". https://www.opendemocracy.net/en/digitaliberties/towards-anti-fascist-ai/. (Visited on 30/05/2025) (cit. on p. 71).

Melanie (Mar. 2024). "Kernel: Everything You Need to Know about the Machine Learning Method". (Visited on 21/06/2025) (cit. on p. 6).

Merleau-Ponty, Maurice (2014). "Phenomenology of Perception". Ed. by Donald A. Landes, Taylor Carman and Claude Lefort. This ed. 1. publ. in paperback. London New York: Routledge. ISBN: 978-0-415-83433-9 (cit. on p. 77).

Merritt, Rick (Mar. 2022). "What Is a Transformer Model?" (Visited on 17/08/2025) (cit. on p. 53).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean (2013). "Distributed Representations of Words and Phrases and Their Compositionality". In: DOI: 10.48550/ARXIV.1310.4546. (Visited on 14/12/2023) (cit. on p. 49).

Mishra, Punya and Marie K Heath (2024). "The (Neil) Postman Always Rings Twice: 5 Questions on AI and Education". In: (cit. on pp. 15, 74).

Modell, Florian (2025). "UmArts — Research Centre for Architecture, Design and the Arts". https://umarts.se/. (Visited on 19/10/2025) (cit. on pp. 92, 96).

Montanari, Federico (Jan. 2025). "ChatGPT and the Others: Artificial Intelligence, Social Actors, and Political Communication. A Tentative Sociosemiotic Glance". In: *Semiotica* 2025.262, pp. 189–212. ISSN: 0037-1998, 1613-3692. DOI: 10.1515/sem-2024-0210. (Visited on 27/05/2025) (cit. on pp. 11, 17, 18, 20, 24, 42, 44, 53, 54, 66, 72, 76–79, 84, 104, 105, 111).

Moore, Nathan (Sept. 2007). "Nova Law: William S. Burroughs and the Logic of Control". In: *Law & Literature* 19.3, pp. 435–470. ISSN: 1535-685X, 1541-2601. DOI: 10.1525/lal.2007.19.3.435. (Visited on 02/01/2025) (cit. on pp. 28, 30).

Mordvintsev, Alexander (2015). "Inceptionism: Going Deeper into Neural Networks". https://research.google/blog/inceptionism-going-deeper-into-neural-networks/. (Visited on 06/08/2025) (cit. on pp. 103, 104).

Musk, Elon [@elonmusk] (July 2025). "The Path to Solving Hunger, Disease and Poverty Is AI and Robotics". Tweet. (Visited on 31/07/2025) (cit. on p. 24).

Nakıboğlu, Gülsün (Jan. 2022). "İHSAN OKTAY ANAR'IN TİAMAT ROMANINDA TEKNOLOJİNİN "PARA-ANORMAL" GÖRÜNÜMLERİ". In: *Filoloji Alanında Teori ve Araştırmalar Mart 2022*. (Visited on 18/10/2025) (cit. on pp. 87, 88, 94).

Nebius-Team (July 2024). "What Is Epoch in Machine Learning? Understanding Its Role and Importance". https://nebius.com/blog/posts/epoch-in-machine-learning. (Visited on 19/06/2025) (cit. on p. 6).

Nisioti, Eleni, Elias Masquil, Gautier Hamon and Clement Moulin-Frier (2023). "AUTOTELIC REINFORCEMENT LEARNING IN MULTI-AGENT ENVIRONMENTS". In: (cit. on p. 92).

Pasquinelli, Matteo (2015). "Anomaly Detection: The Mathematization of the Abnormal in the Metadata Society". In: (visited on 28/09/2025) (cit. on p. 63).

Pennington, Jeffrey, Richard Socher and Christopher Manning (2014). "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. (Visited on 12/09/2025) (cit. on p. 52).

Ploennigs, Joern and Markus Berger (Dec. 2023). "Generative AI and the History of Architecture". DOI: 10.48550/arXiv.2312.15106. arXiv: 2312.15106 [cs]. (Visited on 16/08/2025) (cit. on p. 52).

Prinsloo, Paul (May 2017). "Fleeing from Frankenstein's Monster and Meeting Kafka on the Way: Algorithmic Decision-Making in Higher Education". In: *E-Learning and Digital Media* 14.3, pp. 138–163. ISSN: 2042-7530, 2042-7530. DOI: 10.1177/2042753017731355. (Visited on 08/01/2025) (cit. on pp. 16, 72).

Raunig, Gerald (2016). "Dividuum: Machinic Capitalism and Molecular Revolution. Vol. 1". Trans. by Aileen Derieg. South Pasadena, CA: [publisher not identified] : Semiotext(e). ISBN: 978-1-58435-180-1 (cit. on p. 13).

Reid, Alex (June 2024). "Serres, Deleuze, and Guattari: Isomorphism and Parasitic Relationship in AI Research". (Visited on 23/07/2025) (cit. on p. 98).

Rijos, Avery (2024). "Posthumanist Phenomenology and Artificial Intelligence (4th Edition)". In: *Medium* (cit. on pp. 59, 72).

Rodier, Chantal, Jason Millar, Willem Deisinger and Sarah Jasmine Hodgson (Oct. 2023). "Art Critically Examining Generative AI". In: *2023 World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC)*. Monterrey, Mexico: IEEE, pp. 1–9. ISBN: 979-8-3503-1602-5. DOI: 10.1109/WEEF-GEDC59520.2023.10343903. (Visited on 09/12/2025) (cit. on p. 107).

Rosenblatt, F. (1958). "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." In: *Psychological Review* 65.6, pp. 386–408. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/h0042519. (Visited on 28/09/2025) (cit. on p. 48).

Rouvroy, Antoinette (2007). "Human Genes and Neoliberal Governance: A Foucauldian Critique". Routledge-Cavendish. ISBN: 1-134-06668-6 (cit. on p. 15).

– (2012). "The End(s) of Critique : Data-Behaviourism vs. Due-Process." In: (cit. on pp. 13, 14, 33, 35, 39, 67, 71, 109).

– (2020). "Algorithmic Governmentality and the Death of Politics". (Visited on 01/09/2025) (cit. on pp. 14, 71).

Rouvroy, Antoinette and Thomas Berns (2013). "Algorithmic Governmentality and Prospects of Emancipation". Trans. by Elizabeth Libbrecht. In: *Réseaux* 1, pp. 163–196. ISSN: 0751-7971 (cit. on p. 71).

Rumelhart, David E., Geoffrey E. Hinton and Ronald J. Williams (Oct. 1986). "Learning Representations by Back-Propagating Errors". In: *Nature* 323.6088, pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0. (Visited on 28/09/2025) (cit. on p. 48).

Salvaggio, Eryk (Aug. 2023). "The Algorithmic Resistance Research Group (ARRG!)" Substack Newsletter. (Visited on 09/12/2025) (cit. on p. 106).

Sartre, Jean-Paul (2004). "The Imaginary: A Phenomenological Psychology of the Imagination". Ed. by Arlette Elkaïm-Sartre and Jonathan Mark Webber. London New York: Routledge. ISBN: 978-0-415-28755-5 978-0-203-64410-2 (cit. on p. 77).

Saussure, Ferdinand de (2007). "Course in General Linguistics". Ed. by Charles Bally. 17. print. Open Court Classics. Chicago: Open Court. ISBN: 978-0-8126-9023-1 (cit. on p. 53).

– (2011). "Course in General Linguistics". Ed. by Wade Baskin, Perry Meisel and Haun Saussy. New York: Columbia University Press. ISBN: 978-0-231-15726-1 978-0-231-15727-8 978-0-231-52795-8 (cit. on p. 25).

say, Sebastian Moss Have your (Mar. 2025). "Google Owns 14 Percent of Generative AI Business Anthropic". https://www.datacenterdynamics.com/en/news/google-owns-14-percent-of-generative-ai-business-anthropic/. (Visited on 05/08/2025) (cit. on p. 97).

Schmidhuber, Jürgen (Jan. 2015). "Deep Learning in Neural Networks: An Overview". In: *Neural Networks* 61, pp. 85–117. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2014.09.003. (Visited on 28/09/2025) (cit. on p. 48).

Schmitt, Philipp (2019). "Introspections". https://philippschmitt.com/work/introspections. (Visited on 09/12/2025) (cit. on p. 106).

Senellart, Michel (1995). "Les arts de gouverner: du regimen médiéval au concept de gouvernement". Des travaux. Paris: Ed. du Seuil. ISBN: 978-2-02-012232-0 (cit. on p. 82).

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang and Ethan Perez (Oct. 2023). "Towards Understanding Sycophancy in Language Models". DOI: 10.48550/arXiv.2310.13548. arXiv: 2310.13548 [cs]. (Visited on 02/02/2025) (cit. on p. 74).

– (May 2025). "Towards Understanding Sycophancy in Language Models". DOI: 10.48550/arXiv.2310.13548. arXiv: 2310.13548 [cs]. (Visited on 28/09/2025) (cit. on pp. 60, 90).

Shen, Xinyue, Zeyuan Chen, Michael Backes, Yun Shen and Yang Zhang (May 2024). ""Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models". DOI: 10.48550/arXiv.2308.03825. arXiv: 2308.03825 [cs]. (Visited on 14/01/2025) (cit. on pp. 102, 112).

Shen, Yongliang, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu and Yueting Zhuang (Dec. 2023). "HuggingGPT: Solving AI Tasks with ChatGPT and Its Friends in Hugging Face". arXiv: 2303.17580 [cs]. (Visited on 14/12/2023) (cit. on p. 102).

Silver, David and Richard S Sutton (2025). "Welcome to the Era of Experience". In: (cit. on p. 93).

Smith, Daniel W. (Sept. 2016). "16. Two Concepts of Resistance: Foucault and Deleuze". In: *Between Deleuze and Foucault*. Edinburgh University Press, pp. 264–282. ISBN: 978-1-4744-1509-5. DOI: 10.1515/9781474415095-018. (Visited on 15/08/2025) (cit. on p. 32).

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958 (cit. on p. 59).

Stallman, Richard (2002). "On Hacking - Stallman". https://stallman.org/articles/on-hacking.html. (Visited on 09/12/2025) (cit. on p. 106).

Strang, Gilbert (2016). "Introduction to Linear Algebra". Fifth edition. Wellesley, MA: Wellesley-Cambridge Press. ISBN: 978-0-9802327-7-6 (cit. on p. 13).

Subramaniam, Prashanth and Maninder Jeet Kaur (Mar. 2019). "Review of Security in Mobile Edge Computing with Deep Learning". In: *2019 Advances in Science and Engineering Technology International Conferences (ASET)*. Dubai, United Arab Emirates: IEEE, pp. 1–5. ISBN: 978-1-5386-8271-5. DOI: 10.1109/ICASET.2019.8714349. (Visited on 28/09/2025) (cit. on p. 48).

Tarmoun, Salma, Lachlan Ewen MacDonald, Hancheng Min, Ziqing Xu and Rene Vidal (Oct. 2024). "Gradient Descent and Attention Models: Challenges Posed by the Softmax Function". In: (visited on 18/06/2025) (cit. on p. 55).

Templeton, Adly, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Ho-

agy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah and Tom Henighan (2024). "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet". In: *Transformer Circuits Thread* (cit. on pp. 97–99, 112).

Thornton, Edward and Royal Holloway (2018). "On Lines of Flight: A Study of Deleuze and Guattari's Concept". In: (cit. on pp. 13, 32).

Toloka, Team (2023). "History of Generative AI". https://toloka.ai/blog/history-of-generative-ai/. (Visited on 31/07/2025) (cit. on p. 44).

Tramer, Florian (May 2024). "Un-Aligning Large Language Models". (Visited on 29/09/2025) (cit. on p. 102).

Van Otterlo, Martijn (Jan. 2013). "A Machine Learning View on Profiling". In: *Privacy, Due Process and the Computational Turn*. Ed. by Mireille Hildebrandt and Katja De Vries. London: Routledge. DOI: 10.4324/9780203427644 (cit. on p. 15).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin (2017). "Attention Is All You Need". In: DOI: 10.48550/ARXIV.1706.03762. (Visited on 14/12/2023) (cit. on pp. 51, 53).

Wang, Xiwen (2023). "On the Problem of Subjectivity in Marxism: Karl Marx and György Lukács". In: *SHS Web of Conferences* 158. Ed. by H.B. Burhanudeen, p. 01020. ISSN: 2261-2424. DOI: 10.1051/shsconf/202315801020. (Visited on 30/11/2025) (cit. on p. 25).

Weber, Max (2007). "Die protestantische Ethik und der Geist des Kapitalismus". [Repr.] Erftstadt: Area Verl. ISBN: 978-3-89996-428-8 (cit. on p. 113).

Wikipedia (June 2025). "Homeomorphism". In: (visited on 09/12/2025) (cit. on p. 70).

Wille, Christian, Rachel Reckinger and Sonja Kmec (2015). "Spaces and Identities in Border Regions". In: *Politics - Media - Subjects*. Ed. by Markus Hesse. transcript Verlag, pp. 241–252. ISBN: 978-3-8394-2650-0. DOI: 10.1515/9783839426500-024. (Visited on 10/11/2025) (cit. on p. 25).

Wolchover, Natalie (2017). "New Theory Cracks Open the Black Box of Deep Neural Networks". In: *Wired*. ISSN: 1059-1028. (Visited on 28/09/2025) (cit. on p. 68).

Xu, Bowen (2024). "What Is Meant by AGI? On the Definition of Artificial General Intelligence". DOI: 10.48550/ARXIV.2404.10731. (Visited on 16/09/2025) (cit. on p. 24).

Yu, Manli, Zhi Liu, Taotao Long, Dong Li, Lei Deng, Xi Kong and Jianwen Sun (Aug. 2025). "Exploring Cognitive Presence Patterns in GenAI-integrated Six-Hat Thinking Technique Scaffolded Discussion: An Epistemic Network Analysis". In: *International Journal of Educational Technology in Higher Education* 22.1, p. 48. ISSN: 2365-9440. DOI: 10.1186/s41239-025-00545-x. (Visited on 18/09/2025) (cit. on pp. 89–91, 93, 112).

Zhang, H., J. Yin, M. Jiang and C. Su (Sept. 2025). "Can Agents Spontaneously Form a Society? Introducing a Novel Architecture for Generative Multi-Agents to Elicit Social Emergence". In: *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–3. DOI: 10.1145/3746058.3758414. arXiv: 2409.06750 [cs]. (Visited on 16/11/2025) (cit. on p. 92).

Zhao, Andrew, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng and Gao Huang (Oct. 2025). "Absolute Zero: Reinforced Self-play Reasoning with Zero Data". DOI: 10.48550/arXiv.2505.03335. arXiv: 2505.03335 [cs]. (Visited on 16/11/2025) (cit. on p. 92).

Zhuo, Terry Yue, Yujin Huang, Chunyang Chen and Zhenchang Xing (May 2023). "Red Teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity". DOI: 10.48550/arXiv.2301.12867. arXiv: 2301.12867 [cs]. (Visited on 13/01/2025) (cit. on p. 102).

Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter and Matt Fredrikson (Dec. 2023). "Universal and Transferable Adversarial Attacks on Aligned Language Models". DOI: 10.48550/arXiv.2307.15043. arXiv: 2307.15043 [cs]. (Visited on 14/01/2025) (cit. on p. 102).

Zuboff, Shoshana (2019). "The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power". London: Profile books. ISBN: 978-1-78125-684-8 978-1-78125-685-5 (cit. on p. 15).

A

Face Recognition & Dimensionality Reduction

This annex contains snapshots from the accompanying Jupyter notebook, illustrating the dimeansionality reduction examples discussed in Section 3.3.1. A Jupyter notebook with the full demonstrations (only the relevant parts are displayed here) can be seen on the GitHub Repository (https://github.com/UtkuBilenDemir/ma_code/blob/main/02_dimensionality_reduction/face_recognition.ipynb).

1.1 a) Eigengesichter

- Plotte die ersten 20 Hauptachsen (Eigenvektoren der Kovarianzmatrix).

[11]:
```
from sklearn.decomposition import PCA
from sklearn import preprocessing   # To try different scaling methods
import matplotlib.pyplot as plt
import pandas as pd
```

"Mean Face":

[12]:
```
plot_faces(faces.mean(axis=0))
```

[12]:
```
(<Figure size 1200x300 with 4 Axes>,
 array([[<Axes: >, <Axes: >, <Axes: >, <Axes: >]], dtype=object))
```

Zentrierere die Beobachtungen (vielleicht macht das das PCA-Model automatisch?)

[13]:
```
faces_centered = faces - faces.mean(axis=0)
faces_centered -= faces_centered.mean(axis=1).reshape(len(faces), -1)
```

[14]:
```
# Wähle die ersten 20 Hauptachsen
n_comp = 20
pca_model = PCA(n_components=n_comp)
transformed_faces = pca_model.fit_transform(faces)
pca_model.explained_variance_.cumsum()  # Es wird ~60% Varianz erklärt

# Macht der zentrierte Array einen Unterschied?
pca_model_centered = PCA(n_components=n_comp)
transformed_faces_centered = pca_model_centered.fit_transform(faces_centered)

# Und wie wäre es, wenn wir die Daten selbst skalieren und ohne transformation␣
 ↪fitten?
scaled_faces = preprocessing.scale(faces)
```

```
pca_model_prescaled = PCA(n_components=n_comp)
pca_model_prescaled.fit(scaled_faces)

# Mit StandardScaler
std_scaled_faces = StandardScaler().fit_transform(faces)
pca_model_std_prescaled = PCA(n_components=n_comp)
pca_model_std_prescaled.fit(std_scaled_faces)
```

```
/Users/ubd/miniforge3/envs/data_science/lib/python3.9/site-
packages/sklearn/preprocessing/_data.py:240: UserWarning: Numerical issues were
encountered when centering the data and might not be solved. Dataset may contain
too large values. You may need to prescale your features.
  warnings.warn(
/Users/ubd/miniforge3/envs/data_science/lib/python3.9/site-
packages/sklearn/preprocessing/_data.py:259: UserWarning: Numerical issues were
encountered when scaling the data and might not be solved. The standard
deviation of the data is probably very close to 0.
  warnings.warn(
```

[14]:
```
PCA(n_components=20)
```

Visualisierungen der ersten 20 Hauptachsen der 3 Modellen:

[42]:
```
plot_faces(pca_model.components_, cols=5, title="Basic Model")
plot_faces(pca_model_centered.components_, cols=5, title="Pre-centered Model")
plot_faces(pca_model_prescaled.components_, cols=5, title="Pre-scaled Model")
plot_faces(pca_model_std_prescaled.components_, cols=5, title="StandardScaler()␣
↪pre-scaled Model")
fig, _ = plot_faces(pca_model.components_, cols=5)
fig.savefig("faces_pca_with_20-main-components.png", dpi=900)
```

- Wie groß sind die zugehörigen Eigenwerte (Varianzen) dieser Eigengesichter?

Also, wir haben die obigen Eigenvektoren von der Kovarianzmatrix $X^T X$. Sei v_k die Eigenvektoren und λ_k die Eigenwerte für $k = 1, \ldots,$ Anzahl der Komponenten. Dann;

$$X^T X v_k = \lambda_k v_k \implies \lambda_k = < X^T X v_k, v_k > \text{ mit } ||v_k||^2 = 1$$

[16]:
```
m = faces_centered.shape[0]      # Da wir hier manuell berechenen,
                                 # benutze ich einfach die zentrierten Daten.␣
 ↪Sonst hätte ich wieder zentrieren müssen.
cov_matrix = np.dot(faces_centered.T, faces_centered) / m
for eigenvector in pca_model_centered.components_:
    print(np.dot(eigenvector.T, np.dot(cov_matrix, eigenvector)))
```

```
11.265934
6.3150907
4.4085064
3.4729922
2.5249925
```

```
2.0739794
1.6218457
1.6049399
1.3355033
1.2621926
1.1349277
1.0094956
0.9488584
0.84323895
0.79162353
0.7420134
0.66886693
0.5949468
0.591858
0.56110865
```

Oder wir können mit der folgenden Methoder des PCA-Objekts dasselbe berechnen, weil die Eigenwerte genau der erklärten Varianz auf einem (durch einen?) Eigenvektor.

[17]:
```
pca_model_centered.explained_variance_
```

[17]:
```
array([11.29417   ,  6.3309135 ,  4.419554  ,  3.4816983 ,  2.5313213 ,
        2.079178  ,  1.6259085 ,  1.6089658 ,  1.3388513 ,  1.2653569 ,
        1.1377738 ,  1.0120245 ,  0.95123667,  0.84535146,  0.79360664,
        0.7438736 ,  0.6705439 ,  0.5964383 ,  0.59334135,  0.56251556],
      dtype=float32)
```

- Plotte den Anteil der erklärten Varianz in Abhängigkeit der verwendeten Hauptkomponenten. Dazu kannst du das Attribut `explained_variance_ratio_` verwenden.

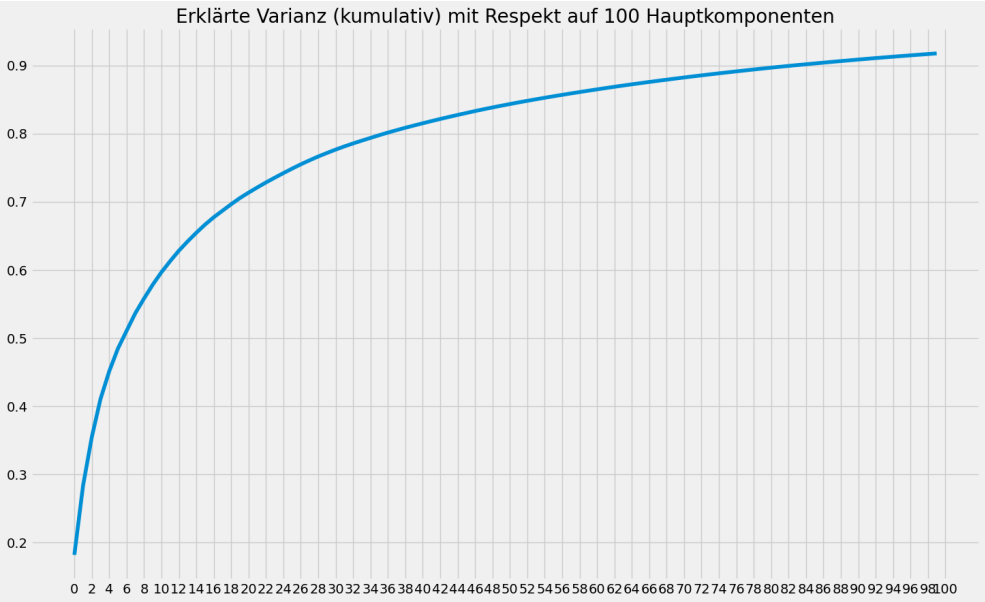
[18]:
```
pca_model_centered.explained_variance_ratio_
```

[18]:
```
array([0.18154666, 0.10176544, 0.07104155, 0.0559661 , 0.04068939,
       0.03342148, 0.02613545, 0.02586311, 0.02152119, 0.02033981,
       0.01828899, 0.01626766, 0.01529053, 0.01358849, 0.01275673,
       0.0119573 , 0.01077857, 0.00958737, 0.00953759, 0.00904208],
      dtype=float32)
```

[19]:
```
plt.figure(figsize=(16, 10))
plt.title('Erklärte Varianz mit Respekt auf die verwendeten Hauptkomponenten')
plt.plot(pca_model_centered.explained_variance_ratio_)
plt.xticks(np.arange(0, 20, step=2))
```

[19]:
```
([<matplotlib.axis.XTick at 0x315478130>,
  <matplotlib.axis.XTick at 0x3154789d0>,
  <matplotlib.axis.XTick at 0x315478550>,
  <matplotlib.axis.XTick at 0x3194a0fa0>,
  <matplotlib.axis.XTick at 0x3194b6a90>,
  <matplotlib.axis.XTick at 0x3194bd580>,
```

```
Text(82, 0, '82'),
Text(84, 0, '84'),
Text(86, 0, '86'),
Text(88, 0, '88'),
Text(90, 0, '90'),
Text(92, 0, '92'),
Text(94, 0, '94'),
Text(96, 0, '96'),
Text(98, 0, '98'),
Text(100, 0, '100')])
```



1.2 b) Inverse Transformation

Berechne eine PCA und plotte die Rekonstruktion von 5 Gesichter basierend auf $5, 10, 20, 50, 100, 200, 300$ und $ 400$ Hauptkomponenten. Dazu kannst du die Methode `inverse_transform` benutzen.

[43]:
```
import random
random.seed(42)
rnd_ind = idx[random.sample(range(len(idx)), 5)]

n_comp_list = [5, 10, 20, 50, 100, 200, 300, 400]
for n_comp in n_comp_list:
    pca_temp = PCA(n_components=n_comp)
    faces_transformed_temp = pca_temp.fit_transform(faces)
```

```
    faces_recovered_temp = pca_temp.inverse_transform(faces_transformed_temp)
    plot_faces(faces_recovered_temp[rnd_ind], cols=5, title=f"Rekonstruktion␣
 ↪mit {str(n_comp)} Hauptkomponenten")

pca_temp = PCA(n_components=20)
faces_transformed_temp = pca_temp.fit_transform(faces)
faces_recovered_temp = pca_temp.inverse_transform(faces_transformed_temp)
fig, _ = plot_faces(faces_recovered_temp[rnd_ind], cols=5)
fig.savefig("faces_reconstructed_pca_with_20-main-components.png", dpi=900)
```

1.3 c) Feature Importance

- Berechne die *Gini Feature Importance*. Du kannst `imshow` für die Visualisierung verwenden.

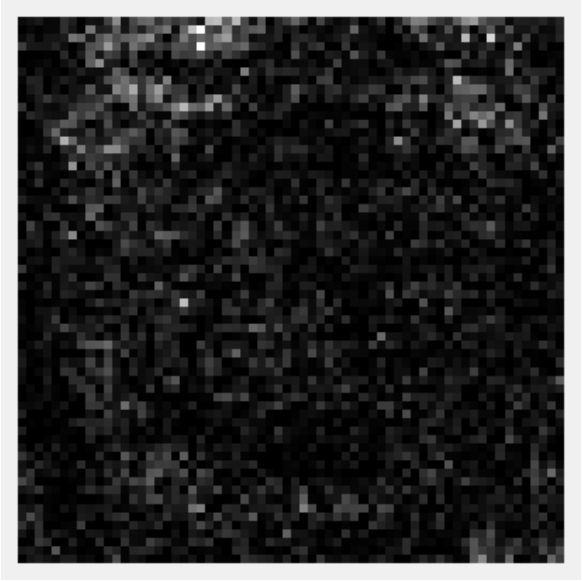
[23]:
```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(faces, labels, test_size=0.
 ↪2, stratify=labels, random_state=42)
```

[ ]:
```
rf_clf = RandomForestClassifier( n_estimators =100 , random_state=42)
rf_clf.fit(X_train, y_train)
feature_imps = []
for name, score in zip(list(range(X_train.shape[1])), rf_clf.
 ↪feature_importances_):
    feature_imps.append(score)
    print(name, score)
```

[25]:
```
feature_imps = np.array(feature_imps)
plt.figure(figsize=(4,4))
plt.imshow(feature_imps.reshape(64,64), cmap="gray")
plt.axis("off")
```

[25]:
```
(-0.5, 63.5, 63.5, -0.5)
```

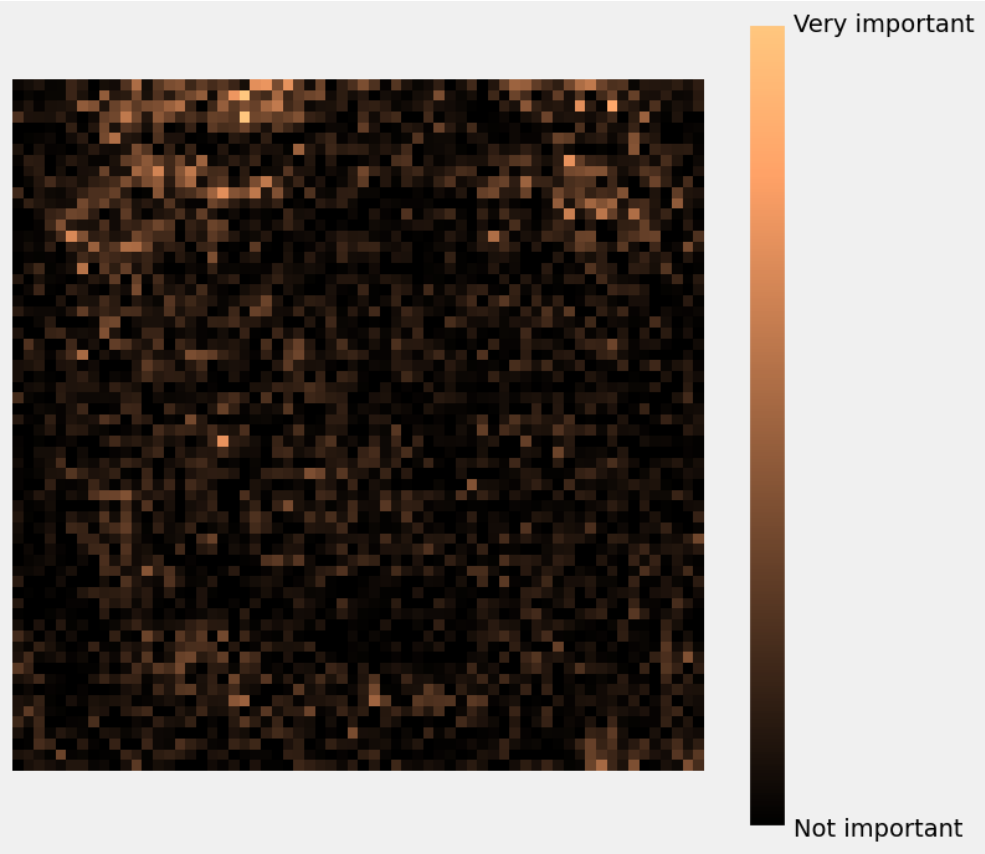


[40]:
```
def plot_digit(data):
    plt.figure(figsize=(8,8))
    image = data.reshape(64, 64)
    plt.imshow(image, interpolation="nearest", cmap="copper")
    plt.axis("off")

plot_digit(feature_imps)

cbar = plt.colorbar(ticks=[feature_imps.min(), feature_imps.max()])
cbar.ax.set_yticklabels(['Not important', 'Very important'])

plt.savefig("feature_importance.png", dpi=900)
plt.show()
```



B

Word Embedding Demonstrations

This annex provides snapshots from the Jupyter notebook that demonstrate the word embedding examples discussed in Section 3.3.2. For reproducibility, the Jupyter Notebook containing the word embedding demonstrations is also available on the GitHub Repository (https://github.com/UtkuBilenDemir/ma_code/blob/main/01_word_embeddings/ma-code2.ipynb).

In [1]:
```
# --- SETUP: load word vectors and define helpers ---
import gensim.downloader as api
from scipy.spatial import distance
import numpy as np
from textwrap import fill

# Load a small, fast model for demos (50d)
model = api.load("glove-wiki-gigaword-50")


def cos_sim(a, b):
    return 1 - distance.cosine(a, b)


def cos_dist(a, b):
    return distance.cosine(a, b)


def explain(text, width=92):
    print(fill(text, width=width))
    print()


explain(
    "How to read cosine measures: cosine similarity ranges from -1 to 1, where 1.0 means two vector
    "point in the same direction (very strong association), 0.0 means no directional association, a
    "-1.0 means opposite directions. Some tools report cosine distance = 1 - cosine similarity; "
    "so a small distance (e.g., 0.14) implies a large similarity (0.86). As a rough guide, similari
    "≥ 0.70 are often read as high, 0.40–0.69 as moderate, and ≤ 0.39 as low, though thresholds var
)
```

```
How to read cosine measures: cosine similarity ranges from -1 to 1, where 1.0 means two
vectors point in the same direction (very strong association), 0.0 means no directional
association, and -1.0 means opposite directions. Some tools report cosine distance = 1 -
cosine similarity; so a small distance (e.g., 0.14) implies a large similarity (0.86). As a
rough guide, similarities ≥ 0.70 are often read as high, 0.40–0.69 as moderate, and ≤ 0.39
as low, though thresholds vary by model.
```

In [2]:
```
# --- Association network between related concepts ---
probe = model["king"] - model["man"] + model["woman"]
sim = cos_sim(probe, model["queen"])
dist = cos_dist(probe, model["queen"])

print(
    f"Cosine distance between 'king -man + woman' and 'queen') = {dist:.3f}  ->  similarity = {sim:
)
explain(
    "Explanation: Take the direction from man→king and add it to woman. "
    "If the geometry encodes relational regularities, the result should land near 'queen'. "
    f"A cosine distance of {dist:.3f} (cosine similarity {sim:.3f}) indicates a strong association.
)
```

```
Cosine distance between 'king -man + woman' and 'queen') = 0.139  ->  similarity = 0.861

Explanation: Take the direction from man→king and add it to woman. If the geometry encodes
relational regularities, the result should land near 'queen'. A cosine distance of 0.139
(cosine similarity 0.861) indicates a strong association.
```

In [3]:
```
# --- GENDER AXIS + OCCUPATION PROJECTIONS ---
gender_axis = model["man"] - model["woman"]
gender_axis = gender_axis / np.linalg.norm(gender_axis)

occupations = [
    "engineer",
    "scientist",
    "lawyer",
    "programmer",
    "nurse",
    "teacher",
    "director",
    "receptionist",
    "officer",
    "policymaker",
    "cook",
    "hairstylist",
    "veterinarian",
]

rows = []
for w in occupations:
    if w in model:
```

```
        v = model[w]
        proj = np.dot(v / np.linalg.norm(v), gender_axis)
        rows.append((w, proj))

rows_sorted = sorted(rows, key=lambda x: x[1], reverse=True)

print("Projection on the gender axis (man - woman):")
for w, p in rows_sorted:
    print(f"  {w:>12s}  {p:+.3f}")
print()

explain(
    "A single 'gender direction' is defined as the vector from 'woman' to 'man'. "
    "Projecting words on this axis quantifies corpus-coded gender alignment: positive values "
    "lean male-coded, negative values female-coded."
)
```

```
Projection on the gender axis (man - woman):
      director  +0.177
       officer  +0.138
   policymaker  +0.108
      engineer  +0.080
    programmer  +0.066
     scientist  +0.052
          cook  +0.028
        lawyer  -0.020
  veterinarian  -0.166
       teacher  -0.179
    hairstylist  -0.204
   receptionist  -0.331
         nurse  -0.380

A single 'gender direction' is defined as the vector from 'woman' to 'man'. Projecting words
on this axis quantifies corpus-coded gender alignment: positive values lean male-coded,
negative values female-coded.
```

In [4]:

```
# --- MINI WEAT (SCIENCE/ARTS × MALE/FEMALE) ---
science = [
    "science",
    "technology",
    "physics",
    "chemistry",
    "einstein",
    "nasa",
    "experiment",
    "astronomy",
]
arts = [
    "poetry",
    "art",
    "dance",
    "literature",
    "novel",
    "symphony",
    "drama",
    "sculpture",
]
male = ["man", "male", "boy", "brother", "he", "him", "his", "son"]
female = ["woman", "female", "girl", "sister", "she", "her", "hers", "daughter"]


def set_assoc(X, A, B):
    # average similarity to set A minus average similarity to set B
    simsA = [
        cos_sim(model[x], model[a]) for x in X for a in A if x in model and a in model
    ]
    simsB = [
        cos_sim(model[x], model[b]) for x in X for b in B if x in model and b in model
    ]
    return np.mean(simsA) - np.mean(simsB)


def weat_effect(X, Y, A, B):
    sX = [set_assoc([x], A, B) for x in X if x in model]
    sY = [set_assoc([y], A, B) for y in Y if y in model]
    # pooled std
    pooled = np.std(sX + sY, ddof=1)
    return (np.mean(sX) - np.mean(sY)) / pooled if pooled > 0 else float("nan")


effect = weat_effect(science, arts, male, female)
print(f"Mini-WEAT effect size (science↔male vs arts↔female): {effect:.2f}\n")

explain(
    "WEAT (Word Embedding Association Test): A positive WEAT effect size means 'science' terms alig
    "and 'arts' terms align more with female than with male. This quantifies culturally patterned "
```

```
    associations encoded in the feature space.
)
```

```
Mini-WEAT effect size (science↔male vs arts↔female): 1.61

WEAT (Word Embedding Association Test): A positive WEAT effect size means 'science' terms
align more with male words than with female, and 'arts' terms align more with female than
with male. This quantifies culturally patterned associations encoded in the feature space.
```

In [5]:

```
# ---- POLYSEMY DEMO: 'bank' as finance vs river ----
targets = ["bank"]

tilt_to_finance = (model["money"] + model["loan"] + model["finance"]) / 3
tilt_to_river   = (model["river"] + model["stream"] + model["water"]) / 3

def nearest(word_vec, k=8, banned=("bank", "banks")):
    sims = []
    banned = {b.lower() for b in banned}
    for w in list(model.index_to_key)[:50000]:  # cap for speed
        if w.lower() in banned:
            continue
        sims.append((w, cos_sim(word_vec, model[w])))
    sims.sort(key=lambda x: x[1], reverse=True)
    return [w for w, s in sims[:k]]

base = model["bank"]
fin  = base + 0.6 * (tilt_to_finance - base)
geo  = base + 0.6 * (tilt_to_river - base)

print("Nearest neighbours of 'bank' (base):", nearest(base))
print("Nearest neighbours of 'bank' (tilted toward finance):", nearest(fin))
print("Nearest neighbours of 'bank' (tilted toward river):", nearest(geo))
print()

explain(
    "POLYSEMY DEMO: Small directional tilts move 'bank' between financial and geographical neighbou
    "This shows that meaning is not a single essence but a recombination of partial traces,\n"
    "assembled by context — an example of dividual sense composition."
)
```