

THE POETICS OF AUTOPOIESIS:
Visual Arts, Autonomy
and Artificial Intelligence



Bruno Caldas Vianna



THE POETICS OF AUTOPOIESIS:
Visual Arts, Autonomy
and Artificial Intelligence

Bruno Caldas Vianna



Contents

| | |
|--|-----|
| Abstract | 7 |
| Tiivistelmä | 10 |
| Abstrakt | 13 |
| Acknowledgements | 17 |
| Introduction | 21 |
| Backgrounds | 27 |
| Generative Visual AI | 75 |
| Artificial General Intelligence | 111 |
| Arrangements around autonomy and language | 135 |
| Aesthetic issues | 167 |
| Conclusion | 183 |
| References | 191 |

Abstract

This thesis project starts an inquiry: is it possible for an artificial entity to create art in a manner completely autonomous from humans? As a practice-based artistic research project, the thesis is guided by the production of five works of art, presented during the study: *Photosynthesis*, *Crank of Life*, *AiKU*, *Post-colonialism* and *Supervision*.

An initial framing of historical efforts to create mechanical artists is followed by an investigation into the development of methods to replicate systems of thought. How can reasoning be represented? Once this challenge is overcome, a path to an artificial thinking machine might appear. This leads us from the development of symbolic logic, starting from Leibniz, onto the works of Charles Babbage and Ada Lovelace, and finally to the inception of the first computing systems as well as the conceptualisation of the Turing machine. It also takes us to art proposals that detach themselves from the materiality of objects towards a conceptualising stance. In concept art, which also appeared in the early 20th century, craft becomes less relevant than idea. I suggest that such artworks can be reduced to instructions like “send a urinal to an exhibition”, a paradigm that would actually be fully adopted and put to practice by later conceptualists like Sol Lewitt and Fluxus.

In the field of computer science, the task of building an artificial mind begins to be tackled in the 1950s. At the time, emerging fields like cybernetics, systems theory, and artificial intelligence are addressing this challenge. However, a fundamental paradox arises: Can an artificial entity truly act autonomously if bound by its creator’s rules? Similar paradoxes arise in various fields: in particular, whether mathematics is self-contained – that is, does it need an external, dogmatic axiom, or can it derive all solutions by using rules that can prove themselves? It is widely accepted that Kurt Gödel proved the first hypothesis, undermining the possibility that an artificially created system can develop autonomous agency. The research looks into two theories that face this dilemma, one most

recent and pragmatic, by Karl Frison, and the more poetic approach of Douglas Hofstadter. The task is clearly beyond the possibilities of this project and remains open. It seems unlikely that an artificial entity will show complete autonomous creativity in the distant future, and in any case it is certainly not happening now. Hence, we turn the research towards understanding the poetics that the efforts towards autonomous creativity can yield. The first artistic component is then presented as a reflection on this potential. *Photosynthesis* is a piece powered by sunlight, with no battery storage or electrical input. Every time it is hit by sun rays, it will draw a silhouette of a tree on a black and white screen, proposing a human-made type of autonomous artistry.

The search for self-organisation continues with the development of the sciences of autonomy and the artistic expressions around them. The paramount concept of *autopoiesis* (by Maturana and Varela) describes biological systems as defined by their boundaries and self-maintenance capacity. Another artistic component, *Crank of Life*, implements John Conway's *Game of Life*, a landmark in complex simulation systems. My version makes the machine's external reliance very explicit, as it requires the user to crank a dynamo to produce the energy to run the simulation.

The breakthroughs of last decade's artificial intelligence revolution, I argue, were fueled exactly by the need to endow it with the ability to understand the world visually, and with the semantic capacity to interpret it. The "solution" to the ImageNet challenge came to cement the dominance of neural networks as the preferred path within AI research, since the project that managed to recognise image categories with accuracy similar to humans was coded with this technology. As a deference to the influence of ImageNet, I propose *Supervision*, a mechanical curation of 12.864 digital images belonging to the Finnish Museum of Photography, powered by the model.

In 2014, the invention of the Generative Adversarial Network (GAN) triggered the explosion of generative visual AI. The plethora of GANs, with their different goals and purposes, is used to introduce a key idea in the thesis, the *Autopoietic Agential Arrangement*. This concept leans on the work of artists who incorporate technology as an essential part of their practice. Many of these pieces

consist of systems that output artwork such as images, text, sound or even performances. My claim is that the most important creative aspect of this practice is not their outputs (which I call the by-products), but the system itself (which I call the arrangement). The search for a *poetics of autopoiesis* could end with the argument that the transcendence of these artworks lies precisely in their potential for creative autonomy and self-maintenance.

However, the invention of *text to image models* waged a profound impact on creative autonomy and culture. The artwork *AIKU*, a haiku generator that takes inspiration from images found on Wikipedia, represents my take on generative AI. The widespread adoption of these systems has societal and environmental consequences: the artistic component *Post-colonialism* is commentary on biases in generative visual neural networks. These models have issues of visual taste bias, and pose challenges to the very concept of copyright and the place of the artistic craft in society.

Yet, the relation between image and written word presents new affordances for autopoietic arrangements. Such processes open new exploratory paths in poetic production. How proficiently do the models synthesise metaphors? How is polysemy represented in the latent creative space of the neural networks? How well can abstract concepts be translated visually?

In conclusion, this thesis proposes the challenge of the construction of an autonomous creative entity to navigate the intricate landscape of artificial creativity. Whilst the project demonstrates that complete autonomy in artistic creativity by artificial entities remains elusive, it unveils a tapestry of potentials in the combined practice of art, technology in general, and artificial intelligence in particular. As technology continues to evolve, this research lays a foundation for the ongoing dialogue about the role of AI in art, the nature of creativity, and the continued quest for poetic artificial autopoietic systems.

Tiivistelmä

Tämä opinnäytetyö haluaa esittää kysymyksen: voiko keinotekoinen entiteetti luoda taidetta täysin ilman ihmisen apua? Tämä opinnäytetyö on käytäntöön perustuva taiteen tutkimushanke ja sitä ohjaavat viisi tutkimuksen aikana esiin nostamaani taideteosta: *Photosynthesis*, *Crank of Life*, *AiKU*, *Post-colonialism* ja *Supervision*.

Aluksi käsittelen mekaanisten taiteilijoiden luomiseen tähtääviä kokeiluja historiassa, minkä jälkeen tutkin ajatusjärjestelmien toisintamiseksi luotujen menetelmien kehittämistä. Miten päättyä voidaan toisintaa? Ajattelukoneen luominen saattaa onnistua, kun tämä haaste on ratkaistu. Tämän kautta käsittelen symbolisen logiikan kehittämistä Leibnizistä Charles Babbagen ja Ada Lovelacen työhön ja lopulta siirryn käsittelemään ensimmäisten tietojenkäsittelyjärjestelmien luomista sekä Turing-koneen konseptointia. Tämä kautta käsittelen myös teoskuvauksia, jotka irtautuvat kohteiden materiaalisuudesta kohti käsitteellisempää ajattelutapaa. Käsitteellisessä taiteessa, joka myöskin sai alkunsa 1900-luvun alussa, taidosta itsessään tulee vähemmän relevantti kuin ideasta. Ehdotan, että tällaiset teokset voidaan pelkistää sellaisiksi käskyiksi kuin ”lähetä urinaali näyttelyyn”, mikä onkin sellainen paradigma, jota myöhemmät käsitteelliset artistit, kuten Sol Lewitt ja Fluxus-taiteilijat, käyttivät ja toteuttivat.

Tietojenkäsittelytieteen alalla keinotekoisien mielen rakentaminen alkoi 1950-luvulla. Tuolloin haaste pyrittiin selvittämään uusien alojen, kuten kybernetiikan, systeemiteorian ja tekoälyn, avulla. Tämä kuitenkin johti perustavanlaatuisen pulman syntyyn: Voiko keinotekoinen entiteetti todella toimia itsenäisesti, jos sitä sitovat sen luoja antamat säännöt? Samankaltaisia paradokseja nähdään muillakin aloilla, erityisesti kysymyksessä siitä, onko matematiikka itsensä perusteleva järjestelmä, eli tarvitseeko se ulkoisen, dogmaattisen perusolettaman vai voiko se päätyä kaikkiin ratkaisuihin käyttämällä sääntöjä, jotka se osaa itse perustella? Kuten tunnettua, Kurt Gödel todisti ensiksi mainitun hypoteesin todeksi, mikä heikensi mahdollisuutta, että keinotekoisesti luotu järjestelmä

voisi kehittyä itsenäiseksi toimijaksi. Tutkimuksessa tarkastellaan kahta tätä ongelmaa käsittelevää teoriaa: yhtä uusimmista eli Karl Frisonin pragmaattista teoriaa sekä Douglas Hofstadterin runollisempaa lähestymistapaa. Ratkaisu on selvästi tämän tutkimuksen mahdollisuuksien ulkopuolella ja jää avoimeksi. Vaikuttaa epätoimennäköiseltä, että keinotekoinen entiteetti pystyisi osoittamaan täydellistä autonomista luovuutta edes hamassa tulevaisuudessa. Ainakaan tällä hetkellä se ei tähän pysty. Siksi tutkimus suuntautuikin erittelemään ilmaisumuotoja, jota autonomisen luovuuden edistäminen voi tuottaa. Tutkimuksen ensimmäinen taiteellinen osa esitelläänkin juuri tämän potentiaalin pohdintana. *Photosynthesis* on auringonvaloa hyödyntävä teos, jossa ei ole akkua tai ulkoista virtalähdettä. Kun auringon säteet osuvat teoksen valokennoon, sen mustavalkoiselle näytölle piirtyy puun siluetti, mikä viittaa ihmistekijän kaltaiseen autonomiseen taiteilijatyyppiin.

Itsenäisen järjestäytymisen etsiminen jatkuu autonomiaa tutkivien tieteiden ja niiden ympärille kehkeytyvien taiteellisten ilmaisujen kehittymisen muodossa. Tärkeä käsite ”autopoiesis” (Maturana ja Varela) kuvaa biologisia järjestelmiä niiden rajojen ja itsenäisen ylläpitokyvyn kautta. Tutkimuksen toinen taiteellinen osa, *Crank of Life*, toteuttaa John Conwayn Game of Life -ajatusleikkiä, joka on monimutkaisten simulointijärjestelmien virstanpylväs. Versioni tekee koneen ulkoisen riippuvuuden erittäin selväksi, koska siinä käyttäjän täytyy vääntää dynamon kampea, jotta hän voi tuottaa energiaa simulaatiota varten.

Uskon, että viime vuosikymmenen tekoälyvallankumouksen läpimurtoja ruokkivat juurikin tarve antaa tekoälylle kyky ymmärtää maailmaa visuaalisesti ja tulkita sitä semanttisesti. ImageNet-haasteen ”ratkaisu” vahvisti neuroverkkojen valta-aseman tekoälytutkimuksen parhaana väylänä, koska projekti, joka onnistui tunnistamaan kuvakategoriat ihmisen tarkkuuden vertaisesti, koodattiin tällä teknologialla. Kunnianosoitukseksi ImageNetin vaikutukselle voidaan nähdä Suomen Valokuvamuseossa esitetty teos *Supervision*. Teos perustuu kyseisen mallin avulla toteutettuun, 12 864 digitaalisen kuvan mekaaniseen kuratointiin.

Vuonna 2014 Generative Adversarial Networkin (GAN) keksiminen käynnisti generatiivisen visuaalisen tekoälyn räjähdysen. Esittelemällä lukuisia eri GAN:eja, joilla kaikilla on eri tavoitteet

ja käyttötarkoitukset, tuon esille opinnäytetyön keskeisen idean eli autopoieettisen toimijuuden luomisen (Autopoietic Agential Arrangement). Käsite nojaa teknologiaa olennaisena työvälteenä käyttävien taiteilijoiden työhön. Monet näistä teoksista koostuvat järjestelmistä, jotka tuottavat esimerkiksi kuvia, tekstiä, ääntä tai jopa esityksiä. Väitän, että tämän käytännön tärkeimpänä luovana näkökohtana eivät ole sen tuotokset (joita kutsun sivutuotteiksi) vaan itse järjestelmä (jota kutsun sovitukseksi viittauksena musiikkiin). Autopoieesin runouden etsiminen voisi päättyä väitteeseen, jonka mukaan näiden taideteosten transzendenssi johtuu nimenomaan niiden luovan autonomian ja itseylläpidon mahdollisuuksista.

Tekstiä kuvaksi muuttavien mallien keksiminen vaikutti kuitenkin merkittävästi luovaan autonomiaan ja kulttuuriin. Haikugeneraattori AIKU, joka saa inspiraationsa Wikipediasta löytyvistä kuvista, edustaa minun näkökulmaani generatiiviseen tekoälyyn. Näiden järjestelmien laajamittaisella käyttöönotolla on yhteiskunnallisia ja ympäristövaikutuksia: taiteellinen osa *Post-colonialism* on kommentti generatiivisiin visuaalisiin neuroverkkoihin kirjautuneista ennakoasenteista. Näihin malleihin liittyy visuaalisen maun vääristymisen ongelmia. Ne asettavat myös haasteita tekijänoikeuden käsitteelle ja taiteellisen työn taidon reunaehdoille yhteiskunnassa.

Kuvan ja kirjoituksen välinen suhde tuo kuitenkin uusia tarjoumia autopoieettisiin sovituksiin. Tällaiset prosessit avaavat uusia tutkimuspolkua runolliseen tuotantoon. Kuinka taitavasti mallit osaavat tulkita metaforia? Miten monimerkityksisyys näkyy hermoverkkojen piilevässä luovassa tilassa? Kuinka hyvin abstraktit käsitteet voidaan kääntää visuaalisesti?

Lopuksi tämä opinnäytetyö muotoilee haasteen luoda autonomia, luova kokonaisuus, jonka avulla voitaisiin sukuloida keino-tekoisen luovuuden monimutkaisessa maailmassa. Vaikka hanke osoittaa, että keino-tekoisten kokonaisuuksien täydellinen autonomia taiteellisessa luovuudessa ei ole vielä toteutunut, se paljastaa potentiaalien kirjon taiteen, teknologian ja tekoälyn yhteisellä kentällä. Teknologian kehittyessä tämä tutkimus luo perustaa jatkuvalla vuoropuhelulle tekoälyn roolista taiteessa, luovuuden luonteesta ja runollisten, keino-tekoisten ja autopoieettisten järjestelmien jatkuvasta etsinnästä.

Abstrakt

Denna doktorsavhandling ger upphov till en frågeställning: är det möjligt för en artificiell entitet att skapa konst helt oberoende av människor? Som ett praktikbaserat konstnärligt forskningsprojekt styrs avhandlingen fem producerade konstverk som presenteras under undersökningen: *Photosynthesis*, *Crank of Life*, *AiKU*, *Post-colonialism* and *Supervision*.

En inledande inramning av historiska försök att skapa mekaniska konstnärer följs av en undersökning om utvecklingen av metoder för att replikera tankesystem. Hur kan resonemang representeras? Så snart denna utmaning är övervunnen kan en väg till en artificiell tänkande maskin se dagens ljus. Detta leder oss från utvecklingen av symbolisk logik, med början hos Leibniz, till arbetena av Charles Babbage och Ada Lovelace, och slutligen till uppkomsten av de första datorsystemen samt konceptualiseringen av Turingmaskinen. Den tar oss också till förslag på konst som tar avstånd från objektens materialitet mot en konceptuell ställning. Inom koncepttillustrering, som också dök upp i början av 1900-talet blir hantverket mindre relevant än idén. Jag menar att sådana konstverk kan reduceras till instruktioner av typen "skicka en urinoar till en utställning", ett paradigmsom senare skulle anammas och omsättas i praktiken av konceptualister som Sol Lewitt och Fluxus.

Inom datavetenskapen började man på 1950-talet att ta sig an uppgiften att bygga ett artificiellt sinne. Vid den tiden tar framväxande områden som cybernetik, systemteori och artificiell intelligens an denna utmaning. Emellertid uppstår det en grundläggande paradox: Kan en artificiell entitet verkligen agera självständigt om den är bunden till sina skapares regler? Liknande paradoxer uppstår inom olika områden, i synnerhet kan man fråga sig om matematiken är självständig, dvs. om den behöver ett externt, dogmatiskt axiom, eller om den kan härleda alla lösningar genom att använda regler som kan bevisa sig själva? Det är allmänt vedertaget att Kurt Gödel bevisade den första hypotesen och underminerade möjligheten att ett artificiellt skapat system kan utveckla autonomt handlande.

Forskningen undersöker två teorier som möter detta dilemma, den senaste och mest pragmatiska, av Karl Frison, och den mer poetiska av Douglas Hofstadter. Uppgiften ligger klart bortom möjligheterna för detta projekt och förblir öppen. Det verkar osannolikt att en artificiell entitet kommer att visa fullständig autonom kreativitet inom en lång framtid, och i vilket fall som helst händer det verkligen inte nu. Därför vänder vi forskningen mot att förstå den poetik som en strävan mot autonom kreativitet kan ge upphov till. Den första konstnärliga delen presenteras sedan som en reflektion över denna potential. *Fotosyntes* är ett verk som drivs av solljus, utan batterilagring eller elektrisk tillförsel. Varje gång den träffas av solstrålar kommer den att rita en silhuett av ett träd på en svartvit skärm, vilket föreslår en mänskligt skapad typ av autonom konstnärlighet.

Sökandet efter självorganisering fortsätter med utvecklingen av vetenskaper inom autonomi och de konstnärliga uttrycken kring dem. Det centrala konceptet *autopoiesis* (av Maturana och Varela) beskriver biologiska system som definieras av sina gränser och sin förmåga till självunderhåll. En annan konstnärlig del, *Crank of Life*, implementerar John Conways Game of Life, en milstolpe inom komplexa simuleringssystem. Min version synliggör maskinens externa beroende mycket tydligt, eftersom den kräver att användaren vevar en dynamo för att producera energi för att köra simuleringen.

Jag hävdar att genombrotten inom artificiell intelligens under förra decenniets revolution drevs just av behovet att ge den förmågan att förstå världen visuellt och med den semantiska förmågan att tolka den. "Lösningen" på ImageNet-utmaningen kom att cementera de neurala nätverkens dominans som den föredragna vägen inom AI-forskning, eftersom projektet som lyckades känna igen bildkategorier med noggrannhet som liknar människors var kodat med denna teknologi. Som en aktning till ImageNets inflytande föreslår jag *Supervision*, en mekanisk kuratering av 12 864 digitala bilder som tillhör Finlands fotografiska museum, drivet av modellen.

År 2014 utlöste uppfinnningen av Generative Adversarial Network (GAN) en explosion av generativ visuell AI. Den stora mängden av GANs med olika mål och syften, används för att introducera en nyckelidé i avhandlingen, *the Autopoietic Agential Arrangement*. Detta koncept bygger på verk av konstnärer som inkorporerar teknologi som en väsentlig del av verksamheten. Många av dessa verk

består av system som genererar konstverk såsom bilder, text, ljud eller till och med framträdanden. Min tes är att den viktigaste kreativa aspekten av denna praxis inte är deras resultat (som jag kallar biprodukter) utan själva systemet (som jag kallar arrangemanget). Sökandet efter en *poetics of autopoiesis* kan sluta med argumentet att transcendensen hos dessa konstverk ligger just i deras potential för kreativ autonomi och självunderhåll.

Uppfinningen av text-till-bild-modeller utgjorde dock en djupgående inverkan på den kreativa autonomin och kulturen. Konstverket *AIKU*, en haiku-generator som hämtar inspiration från bilder på Wikipedia, representerar mitt perspektiv på generativ AI. Den breda användningen av dessa system har samhällliga och miljömässiga konsekvenser: den konstnärliga delen *Post-colonialism* är en reflektion om partiskhet i generativa visuella neurala nätverk. Dessa modeller har problem med visuell partiskhet när det gäller smak och ställer utmaningar för själva konceptet med upphovsrätt och det konstnärliga hantverkets plats i samhället.

Ändå presenterar förhållandet mellan bild och skrivet ord nya möjligheter för autopoietiska arrangemang. Sådana processer öppnar nya utforskande vägar inom poetisk produktion. Hur effektivt kan modellerna syntetisera metaforer? Hur representeras polysemi i de neurala nätverkens latent kreativa utrymme? Hur väl kan abstrakta begrepp översättas visuellt?

Sammanfattningsvis föreslår denna avhandling att det finns en utmaning att konstruera en autonom kreativ entitet för att navigera i det komplicerade landskapet av artificiell kreativitet. Även om projektet visar att fullständig autonomi inom konstnärlig kreativitet genom konstgjorda entiteter fortfarande är svårt att beskriva, avslöjar det en mångfald av potentialer i den kombinerade praktiken av konst, teknologi i allmänhet och artificiell intelligens i synnerhet. Eftersom teknologin fortsätter att utvecklas lägger denna forskning grunden för en fortlöpande dialog om AI:s roll i konsten, kreativitetens natur och den fortsatta strävan efter poetiska artificiella autopoietiska system.

ACKNOWLEDGEMENTS

In 2020, on the scorched global landscape of the pandemic, Martin and Mariela left with me for one year of adventures on the edge of the Arctic circle, in a country very different in climate, culture and costumes. What at first seemed to be an escape from harsh restrictions of movement and conviviality that prevailed in our city, turned out slowly to be a much stricter confinement. Even classes and intellectual exchanges offered by a university environment became unavailable. At the same time when museums, libraries and restaurants started to close, a tough winter of temperatures below -20 degrees Celsius and grey days fell upon us, limiting our living space to the bounds of a remote student apartment. Yet Martin and Mariela endured the whole ordeal without a complaint; in fact, if it weren't for their constant presence of love and affection, it could have become a miserable experience. To them, I will always be thankful. To them, as always, all my love.

The project wouldn't have been possible without the support of the University of the Arts Helsinki, mainly in the form of a two-year employment contract. Without it, it would have been impossible to dedicate myself and finish the degree in the timeframe proposed. I also appreciate the support for attending conferences, which were essential to shape my view. Uniarts was fantastic in its ability to provide a place so welcoming for experimental research and enquiry. Likewise, the support of the Finnish Cultural Foundation (SKR) in the form of a two-year grant was fundamental in the last phase of the project.

I appreciate the external supervision from Roc Pàres, who often contributed with critical analysis and provoking reflections on the work, including the provision of a view on copyrights of artistic works in the opposite field to mine. It allowed the construction of a rich debate on the mechanics of creation and economic survival – which I also see as a particular kind of autonomy. I would also like to thank Samir Bhowmik for reading the manuscript at an early point, with valuable insights. Rodrigo Nunes got me hooked

on thinking about cybernetics as a tool for understanding social organisation in classes at the Catholic University in Rio de Janeiro 10 years ago. Much before that, João Luiz Vieira made me appreciate the difficulties in translating concepts on the big screen in my undergraduate years.

Finally, my sincere thanks to Mika Elo for his balanced, assuring and always supportive supervision of the whole process. Thanks for believing in the project and for accepting my idiosyncrasies, which carried on into the manuscript. There are many things that I could probably do differently if I were to start again — but I could never pick a better supervisor.

INTRODUCTION

Can a non-human entity autonomously create art? A foundational motivation for this curiosity was the development, around 2018, of several machine learning systems that allowed for the generation of original and unique two-dimensional visual artworks that could replicate any given artistic style.

Revisiting the research proposal, I realise how short I fell from using some of the theoretical tools I intended to use — complexity, self-emergence, and their foundations in thermodynamics. When faced with the actual mechanisms used in artificial intelligence, the investigation had to deviate into the historical and current frameworks that are invested in creating autonomous entities. On the other hand, it is satisfying to see that the research question only became more relevant in the years that the doctorate took place. Finding the *poetics of autopoiesis* in artificial generative systems became an increasingly more rewarding task, in that the configurations of such systems became more complex, incorporating exquisite linguistic features and challenging societal issues.

It is not far-fetched to state that this research was overrun by the vertiginous evolution of generative artificial intelligence in these four years. I can only hope this can justify the fact that the manuscript spans topics so varied. In the first chapter, *Backgrounds*, I make a necessary incursion into the history of AI. I delineate the historical precedents for building artificial creative contraptions — which end up embracing a quest for a symbolic representation of thought. These methods, in their turn, developed into the appearance of computing machines, and finally into the very concept of artificial intelligence. This account is created with foremost attention to visual manifestations. Throughout the chronology, the entanglements between automation, autonomy and art are also documented.

Chapter two, entitled *Generative Visual AI*, looks into the main landmarks in the last few decades of evolution of computer vision and generative visual AI. This led me to introduce a seminal concept discovered in the fabric of the research, the *autopoietic aesthetic*

arrangement. The chapter finishes by introducing the ground-breaking linguistic turn of artificial generative visual systems.

Chapter three, *Artificial General Intelligence*, faces the difficult subject of an artificial general intellect, a construct that presupposes the emergence of entities with a cognition comparable to humans. The importance of the subject stems, naturally, from this being an insurmountable requirement to envision an autonomous artist on a par with its human counterparts. Orbiting concepts such as surprise, cognition, and agency are also presented.

The subsequent chapter, *Arrangements around Autonomy and Language*, looks into experiments around autonomous artists, and delineates my own explorations in language, meaning and the latent space of image generation prompt from text. An enquiry into some tools of semiotics and the philosophy of language reveals that these theories deserve careful revisiting in the face of the new configurations offered by language-based models. In *Aesthetic Issues* I introduce findings around the training of such models, as well as implications for copyright and the role of human artists as special illustrators.

It is paramount to note that all of this is done in a context of artistic research. For a good overview of the potential and issues of such a framework, Wilson and van Ruiten's *Handbook for Artistic Research Education* is a good reference (Ruiten and Wilson 2013). Henk Borgdorff proposed three relations between art and research (Borgdorff 2006). 'Research on the arts' sees art as an object to be studied from a distance. 'Research for the arts' would be the applied investigation that results in new methods and tools, such as 'particular alloys used in casting metal sculptures'. Finally, 'research in the arts' would be the approach where 'the artistic practice itself is an essential component of both the research process and the research result.' The separation between subject and object is obliterated; the artwork informs the research and is informed by the author's curiosity. The works described in this thesis might sometimes serve the second definition, by suggesting and sharing methods to incorporate autonomous practices into generative systems. But their ultimate goal is to be incorporated into the thesis as the evidence that demonstrates a concept, a subjective but informative epistemological component.

The descriptions of the artworks are admittedly succinct. As an artist, I have always found the requirement of an explanation for the artwork troublesome, and I consider it better practice not to delimit interpretations and reactions of viewers by providing an accompanying analysis. I tried my best to build arguments around the pieces as much as I found indispensable, considering they are presented as research evidence in a doctoral thesis. But I still left enough room for readers to make their judgements. While paradoxical, this stance can also explain the fact that the artwork I consider the most powerful, *Photosynthesis*, is the one with the shortest description.

As such, theoretical delvings are accompanied by a hands-on creative approach to both historical and state-of-the-art applications of the theory. These artistic components manifested themselves mainly in two events: the *Singular* exhibition of the 2021 Uniarts Helsinki Research Pavilion, which took place at the Hietsu Pavilion in Helsinki, and the 2022 *Paradoxes of Photography* exhibition, held at the Finnish Museum of Photography in Helsinki. In the first exhibition, the works *Photosynthesis*, *Crank of Life*, *Alku* and *Post-Colonialism* were displayed. A fifth work, *Slaughterhouse*, was part of the show but was not included in the thesis frame. It consisted of AI-generated images that, through the exploration of the latent space, depicted eerie chimeras, in-between creatures, half-human half-animal. Looking backwards, I realised they were more of a curiosity, certainly a secondary effect of the ability of neural networks, not exclusive of them. I didn't think they added new findings to the research.

The latter exhibition brought the piece *Supervision*, the last of the artistic components belonging to the project. Even if the artworks pertaining to the thesis are documented in the research catalogue, their textual descriptions and relations to the research are woven into the manuscript, where I try to situate them within a historical and motivational context.

The outcome of this manuscript is inevitably affected and informed by my own background experience. Since my undergraduate years, I have played a juggling act consisting of the manipulation of visual narratives, moving images, interaction and automation. During the years leading to my degree in filmmaking, I also attended a computer engineering college, which was interrupted by the opportunity to pursue a master's degree where I was able to

combine art and technology. In the almost two decades between this last degree and the return to academia with the current doctorate, I pursued a variety of projects. Some of these explored exactly the limits of automation in narrative and artmaking, as well as the intricacies of trying to convey concepts and emotions to the audience through bidimensional images. This clash between poetics, aesthetics and *autopoiesis* continues to instigate my work, and its debris makes up much of the findings of this thesis.

This is a good place to clarify my use of these terms, two of which are in fact part of the title. *Autopoiesis* is borrowed from systems theory, and its conception is described in more detail in the *Background* chapter. To be succinct, in my view it represents the capacity of a system to *thrive* on its own, more than just to survive and reproduce. This thriving is used here in a poetic sense. But what does it mean to be poetic?

A classical definition of *poetics* is that of an Aristotelian science that tries to understand the process through which certain texts provoke a mesmerising effect on its readers, a result that transcends the immediate meaning of the words they carry. Applying this description to the title is a fair approach: I am trying to comprehend how autonomous entities can elicit a sense of awe, even if they use non-textual supports. In the sections that deal with the linguistic affordances of text-to-image models, the term harks back to a more literal application, as it is used to analyse the effects of written words.

Aesthetics, while not explicitly in the title, is connected to the expression *visual arts*. The reason to not employ it is that *aesthetics* has been used colloquially to refer to much more than fine arts, and is sometimes even a plain synonym for *beauty*. Alexander Baumgarten originally appropriated this Greek term to describe a science of cognition, ‘the theory of the liberal arts, the logic of the lower capacities of cognition [gnoseologia inferior], the art of thinking beautifully, the art of the analogon rationis’ (Baumgarten 1750). This original definition is semantically very close to what I understand as *poetics*. But I believe that its most common connotation now refers to the set of traits within an artwork that provoke the *poetic* effect described before, and this is the meaning that prevails in this manuscript.

The adjectives *poetic* and *aesthetic* follow a similar logic in the thesis. I generally say something is *poetic* when I mean it possesses some unexplainable transcendence (which *poetics* try to explain). And I will qualify as *aesthetic* artwork that has a set of features and qualities to this effect. I tried my best to denote the different applications of these terms throughout the test, but some overlaps might remain.

Given my own background and the impossibility to face a study of art in itself, there is a conscious decision to focus on visual, two-dimensional, non-time-based artwork. I have mostly forsaken incursions into music, sounds, performing arts and other approaches. This is a somewhat contradictory statement, since my practice (and even the art pieces that comprise the thesis) go beyond the flat surface of pictures to propose multidimensional and multisensorial contraptions. I stand by my choice of words, since even the most sculptoric of these works still have surface images and texts as their main focus. Images here serve as case studies that inform the research on universal traits of creativity. In the course of the arguments, I sometimes trespass beyond these restraints to convey a discourse on the act of creation in general, such as when I delineate the concept of *autopoietic aesthetic arrangement*.

This is a project that on several levels was only able to scratch the surface of the issues at hand. The dismissal of the original theoretical framework around emergence says nothing about its usefulness to tackle the subject. On the contrary, the investigations on semantic imaging and autopoietic arrangements point to a huge potential for exploratory production based on complexity theories. The difficulty in relating philosophies of language to the text-to-image models only shows that there is much more work to be done. The explorations of semiotics in latent space have just begun. And finally, the fact that the evolution of machine learning does not seem to be slowing down in the near future makes such probes imperative.

BACKGROUNDS

From verb to thing, command to action. The history of technology is also a history of achieving automated means of production, to ease the burden of performing tasks necessary for survival. But surviving is not the only goal that defines us as humans: we also want to belong, leave traces, enjoy and appreciate our particular interpretations of beauty. And these tasks can also be made easier or more sophisticated by machines.

In this section, I will look into the background of artificial intelligence tools that support the development of generative arts. While there is a long history of man-made contraptions to support the creation of fine arts, I will focus of course on the resources made possible by computer science, in particular the field that is known as artificial intelligence.

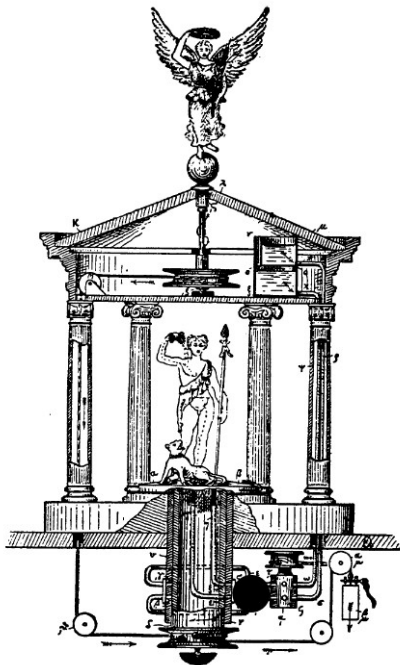
Automation in antiquity

As the science of creating machines that learn and create from data becomes more and more complex, artificial intelligence is becoming increasingly difficult to define. For most of the subjects covered in this work, the concepts of machine learning or deep learning could also be used. No matter how we denote these concepts, their current development is so fast that their naming could be obsolete in a year or two: from the beginning of this research in 2019 until the time of its publication, the focus and breadth of AI systems dedicated to the production of graphical material has changed considerably.

It could be stated, in any case, that the idea of simulating the workings of human imagination and reasoning started with the idea of creating machines that can make calculations. Examples of devices that help with arithmetic tasks have existed since prehistory. The Incas used a calculating device named Yupana, which consisted of a clay table with different slots where grains of different colours were used. Only a few examples of Yupanas were found, but one early account of the cultures of the Andes depicts one in its cover. It was

written around 1615 by the indigenous Peruvian Felipe Guaman Poma de Ayala. Unfortunately, the writer did not provide details of its workings, and the method used for operations performed on this board is still a matter of discussion (Moscovich 2016). Another device named Quipu was used by the Incas to store numbers on knots made on several strings sewn together (Leonard, Shakiban and Yupana n.d.). In the West, the abacus was widely adopted since its appearance more than 4,000 years ago, even reaching China and finally Japan in the 14th century of our age (Gullberg 1997). Mechanisms meant to measure position and astronomical information were devised in ancient Greece, like the astrolabe and the antikhytera mechanism (Freeth et al. 2006), but the versatility of biological brains could not be replicated until programmable computers came about. An early example of a programmable machine is also one of the first documented inventions meant to be used in the arts.

Hero of Alexandria was an engineer, inventor and geometrist who lived around the first century CE in Egypt. Although very little about his personal life is known — not even approximate dates of birth and death — he left very detailed volumes about his creations.



Hero's theatre automaton.
From (Xagoraris and Piene 1991).

These include pneumatic mechanisms, measuring devices and war machines. In a book aptly named *Automatopoietica*, Hero proposes a mechanical theatre. Figurines — sculptures — of the characters move on a small stage, activated by a system of gears and pulleys powered by dropping weights (Xagoraris and Piene 1991).

Although there is no sign of what would be called intelligence here, what is relevant for this research is the emergence of a programmed machine. A series of actions are performed automatically, following a score. The method is similar to what generative art algorithms do, even if the results are quite different: Hero expected the performance to be at the same every time, while computer artists of our times hope for unexpected output through the use of randomness and artificial intelligence. Most importantly, the programming of the play could not be changed. Such automation would also appear in the ‘Book of Ingenious Devices’ published in 850, by three brothers known as Banu Musa, who were working for Abbasid Caliph in Baghdad, in present-day Iraq (Banū-Mūsā, Hill, and Banū-Mūsā 1979). They describe devices such as automatic water fountains and mechanical musical instruments. One of them, a water-powered flute, could be considered the first programmable device, since different melodies would be produced by the use of different rotating drums. Another manuscript with such a device would only appear in the 14th century in Catalonia, a musical carillon controlled by a crank. Later in the same century, the development of automatic carillons became common in Europe, and many of them could be programmed (Koetsier 2001). Again, the early origins of automation are associated with the arts, this time with the execution of a musical piece.

Artistic and symbolic contrivances

Understanding where our fascination with automation comes from could easily be the subject of a whole investigation of its own. But one of its motivations could be the perspective of a workless society, where we wouldn’t have to strive for necessities like nourishment, housing, entertainment. In a utopian perspective, machines could take care of this and leave humans ‘free’ of the burdens of survival and finally be ready to enjoy a life without responsibilities.

The notion is challenged when looking at nature: can the word effort be used to define the growth of a tree, nourished by nutrients and sunlight? How about pre-agricultural societies, living off what can be gathered?

The idea of automation is so enticing that a fake chess-playing machine made in the late 18th century became enormously famous, being presented at several fairs and courts. The Automaton chess player was a model of a human head and torso seated in front of a chess board, built by Wolfgang von Kempelen in 1770. It became known as the Mechanical Turk, as the figure wore a turban and portrayed stereotypical Middle Eastern traits. The model was actually controlled by a human hiding inside the table; an intricate system of sliding seats allowed the presenter to open several doors that would reveal fake mechanisms while hiding the actual operator (Standage 2003). This was a time when such machines abounded, motivated by advances in clockwork machinery. The models made by Pierre Jaquet-Droz, a watchmaker from Neuchâtel, Switzerland, were particularly complex. One of them could write messages of up to 40 characters, stored on a letter dial. It can be considered one of the earliest examples of mechanical memory. Jaquet-Droz also made music-playing devices, like automated boy figures playing the flute (Nahum 2017).

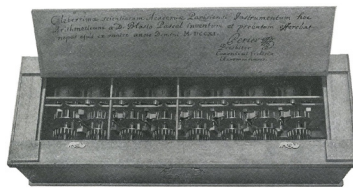
But the idea of an artificial chess player is so seductive that it persisted through the centuries in different ways. On one hand, the challenge of having a machine beat a human at chess was one of the inspiring goals for computer science. This was achieved in 1997 when Deep Blue, a system comprised of special hardware and software designed by IBM, beat the human chess champion Garry Kasparov (D. Clark 1997). Ironically, the name Mechanical Turk was adopted by Amazon when they created a service for hiring remotely located humans to perform massive distributed tasks that computers could do. This product (and its competitors) became one of the driving forces behind the giant leaps in machine learning that occurred in the 2010s.

While the fake chess player was touring around Europe, efforts were being made to build actual computing machines. Making mathematical computations was always seen as one of those tasks we wished would become automated, and several contraptions were

invented to facilitate these chores. Scottish mathematician John Napier published a multiplying machine based on rotating rods in 1617, which helped him publish logarithm tables. Blaise Pascal came up with a design in 1642 that became known as *Pascaline* and could do additions and subtractions. In 1804, Joseph Maria Jacquard invented a loom that could be programmed: different patterns of weaving would be determined by the use of punched cards. This input method would become a standard for machines, including electronic computers, almost into the 21st century.



Napier rods (by Stephen Dickson, CC-BY-SA-4.0)



Pascaline (Public domain)



Jacquard loom (by George Williams, CC-BY-SA-3.0)

A design for a multipurpose programmable engine finally appeared in England in the 19th century. Mathematician and inventor Charles Babbage had first conceived of a mechanical calculator (the Difference Engine) meant to produce tables of logarithms (like Napier's) and trigonometry functions. While this engine was being built, he realised he could design a machine that could be programmed to perform any calculation (Collier, Bruce 1998).

This project incorporated many features that can be found in today's computers. It could store about 1,000 40-digit numbers. This data could be transferred to the 'mill', or what would today be called a central processing unit (CPU), and could perform different operations using the supplied numbers. It could also control the flow of data with logical tests — today's programming languages' if-then statements. Three different models of punch cards were used

for inputting constants, arithmetical functions and memory operations. Babbage wrote several procedures for his device, including programs to treat polynomials and find Bernoulli numbers (Bromley 2000; Bromley 1982).

Babbage's designs were not finished during his lifetime. In fact, a working Difference Engine specimen was only completed in 1992, while the Analytical Engine remains unedited today. It was, in any case, the first concept of a programmable machine, and as such it allowed for the appearance of who is considered the first programmer, Ada Lovelace. This young mathematician became aware of Babbage's devices through her tutor and friend Mary Somerville. Fascinated by the invention, Ada translated an article about it from Italian to English and added her own notes, which are three times as long as the original article. In one of these notes, labelled 'G', she describes a method to generate Bernoulli numbers — a difficult-to-compute series based on the powers of integer numbers — using the engine, which is considered to be the first program ever written. The claim is disputed, since Babbage himself had previously written similar methods for his machine. However, in the same note, almost in passing Lovelace makes an observation that became known as the 'Lovelace Objection' against artificial intelligence:

'The Analytical Engine has no pretensions whatever to *originate* anything. It can do *whatever we know how to order it* to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths.' (Lovelace 1843, 666-731)

Ada Lovelace had a background of interest in arts and poetry. Her concern in speculating on the ability of such a machine to *originate* new works is understandable: she was the daughter of the poet Lord Byron, and even though her father abandoned the family months after her birth, she was used to the company of writers and scientists (Lovelace and Toole 1998). The discussion about Lovelace's objection is taken up by several scholars; it is also central to this thesis and will be tackled later on. It is worth noting that the notion of *to originate* is problematic even for humans, since we also rely on the weight of culture and previous achievement to produce new things: there is no completely original 'origination'.

The will to originate is related to *agency*, a hard to delimit, multi-dimensional concept. For now, to the effect of discussing Lovelace's notes, let us define it as a will to create, common to both artists and inventors. It might be the closest interpretation of the mathematician's concerns, since machines would not take the initiative to produce anything unless programmed to do so. And finally, there is also the notion that machines should not even be compared to humans. If we consider them as having their own particular cognition, they originate things in their own ways, some of which are out of reach for biological-based beings.

Other contraptions that were meant to perform calculations were planned and built after Babbage. Percy Ludgate in Dublin, Leonardo Torres y Quevedo in Madrid, and Vannevar Bush in Boston, developed independent models of analogue computers in the early 20th century (Randell 1982). But the biggest push towards calculating machines came in the Second World War and the need to decode cryptographic messages. In England, a facility was established in a country home named Bletchey Park, with the main purpose of supporting military intelligence. This centre operated machines named *bombes*, hard-wired calculators that facilitated the codebreaking of German Enigma machines. The Colussus machine, often considered the first programmable electronic computer, started operating from there in early 1944 (Copeland 2006).

Representing reasoning

In parallel to the history of hardware for automation and thought, and certainly no less important, is the evolution of formal systems meant to create abstractions for reasoning. These developments are a very interesting side companion to the history of mathematics and philosophy itself. To create a machine that thinks, a method that replicates human processes of deduction and reckoning must be developed. This was appreciated simultaneously in different cultures in the first millennium BC. Aristotle's work *Prior Analytics*, thought to have been written around 350 BC, describes what he named *sylogisms*: 'All men are mortal. If Socrates is a man, then Socrates is mortal' or 'If every A belongs to B and every B belongs to C, then every A belongs to C'. In the Indian philosophy system

Nyanya, described as early as 450 BCE, inference is one of the paths to knowledge (Nyaya | Indian philosophy | Britannica n.d.). Also in India, a Sanskrit grammar, the *Aṣṭādhyāyī*, was written by scholar Pāṇini in the sixth century BCE that provided thousands of rules and used recursion and metarules (Cardona 1997). The Mohist tradition in ancient China, which was very influential during the Warring States Period, developed concepts of contradiction and expression evaluation (Willman 2021).

Middle Eastern tradition greatly expanded the works of Aristotle. Alfarabi (born in 950), Avicenna (born in 1037) and Averroes (born in 1198) criticised and adapted Greek logic to Islamic principles. Al-Khwarizmi was a mathematician working in Baghdad in the eighth century CE, whose Westernised name gave origin to the term algorithm. The title of his book *Hisab al-jabr w'al-muqabala* named the science of algebra. Different step-by-step methods for solving mathematical tasks are described in this treaty. Similar methods, however, had already been described as early as in Babylonian times, in clay tables dating to 1800–1600 BCE (Knuth 1972).

The development of reasoning systems continued sparsely in the West. Mallorcan philosopher Ramon Llull (1232–1315) devised combinatory logical systems for the production of knowledge, where letters and diagrams were matched to provide answers to questions (Eco 2017). Llull's work was read by Gottfried Leibniz (1646–1716), who believed that human thinking could be expressed by a system analogue to mathematical formulas:

“The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say: Let us calculate [*calculemus*], without further ado, to see who is right’ (in *The Art of Discovery*) (Leibniz and Wiener 1951, 51)

Leibniz worked on formal logic, defining concepts like sets, empty sets and negation. However, his writings on the subject remained unpublished until the 1840s, when modern foundations of logic were already being settled by George Boole in his *Mathematical Analysis of Logic* (1847). Boole's symbolic logic — which was rectified in a later

work in 1857 — became known as Boolean algebra, which forms the basis of modern electronic circuit design and computer programming languages. At the core of this algebra is the manipulation of binary symbols — 0 and 1 , or *false* and *true*. The basic operations of these symbols are conjunction (AND), disjunction (OR) and negation (NOT). Conjunction is defined as returning a *true* value if both inputs are *true*, for instance. Negation returns the opposite value — 0 when 1 is the input, *true* if the input is *false*.

A secondary Boolean operation known as exclusive disjunction (XOR) became the pinnacle of a discussion regarding the potential of neural networks in the early 1970s. XOR is defined by outputting *true* if one and only one of the conditions is *true*, and *false* if both conditions are *true* or *false*.

| x | y | x AND y | x OR y | x XOR y |
|---|---|---------|--------|---------|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 | 0 |

The research on a symbolic system for reasoning continued, in parallel, in the work of Gottlob Frege (1848–1925). In the book *Begriffsschrift* (1879), the German philosopher lays out a formal system by the same name as the title, which can be translated as *concept writing*; in other words, a notation system for ideas.

Charles Sanders Peirce, an American mathematician and philosopher, also worked on the foundation of mathematical logic for symbols. He expanded on Boole’s work (Peirce 1865) and developed much of the formalisation of logic simultaneously and unaware of Frege’s efforts. But Peirce also speculated on the nature of thought itself: how are new facts deduced from known ones? How can new inferences be made from a heterogeneous corpus of knowledge? Peirce formulated the concept of *abductive reasoning*, which is still used when trying to get machines that imitate human thought (Peirce 2019). I will briefly go over the different types of reasoning as they are important for the study of artificial intelligence.

Deductive reasoning can be called the most ‘logical’, and was already described by Aristotle:

All X are Y .
All Y are Z .
Therefore, all X are Z .

Or:

All wizards can fly.
Harry is a wizard.
Therefore, Harry can fly.

Inductive reasoning allows us to create reasonable deductions which are likely to be true, but not guaranteed. For instance, a pattern is observed on a large sample of the population, and we infer that it applies to the whole population. Suppose I visited a new place for a week, and I noticed that birds sang every day at sunrise. Therefore, my conclusion is that the birds in this place always sing at sunrise. But my observation cannot guarantee that.

To show what Peirce defined as abductive reasoning, what is needed is a body of evidence that might explain a phenomenon. Again, this can't be used as a proof — other explanations might also be valid. Abduction is used extensively in the medical field, when a set of symptoms adds up to a diagnosis that might not be conclusive. It is a type of reasoning that relies on accumulating knowledge and experience, whether it be systematically or intuitively organised.

Another important piece of Peirce's work is an axiomatisation of natural numbers; that is, creating a set of principles that cannot be proved but that will work as a foundation upon which a complete mathematical science can be built.

The research on symbolic reasoning continued in the late 19th century, as it became a somewhat popular topic. Lewis Carroll, notable for the *Alice in Wonderland* children's book series, was also a mathematician concerned with describing symbolic logic in a way that would be accessible to the general public. His *Symbolic Logic* volumes (1896 and 1977) were written with this goal, but also proposed original methods such as the *truth tree*, a procedural method for propositional logic.

In the late 19th century, Italian logicist Giuseppe Peano formalised some basic axioms, in particular the natural numbers (Peano 1889). Peirce had also proposed a method with the same goal (Peirce

1881) but the Italian's approach is what became the standard in modern mathematics. As the corpus of mathematical systems for reasoning grew, it made sense to attempt to summarise and systematise these findings. Alfred Whitehead and Bertrand Russell's *Principia Mathematica*, a three-volume book published first between 1910 and 1913, was an effort in that sense. *Principia* was an ambitious endeavour that aimed to prove that all mathematical propositions could be deduced from a small set of principles — the axioms provided, for instance, by Peano. Being axioms, Russell naturally does not claim that the small set can be proved: '[...] the explanation of the fundamental concepts which mathematics accepts as indefinable [...] is a purely philosophical task.' On the contrary, these formulations must be accepted without proof: they are the scaffolding on which the building of mathematics can be erected. The *Principia* became a landmark in the development of formal mathematics, supporting crucial computational works from Turing and others, even if flaws might have been detected in its theory (Linsky and Irvine 2022).

One of the greatest consequences of their effort was the questioning of the idea of trying to prove the whole of mathematical theory. Is anything being proved, if in order to do that we must start with unprovable axioms? This was one of the questions raised by the German scholar David Hilbert in a speech in 1900 in Paris, where he described ten problems that were later published as part of a list of twenty-three unresolved problems in mathematics (Hilbert 1902). In what became known as Hilbert's second problem, he asks:

'When we are engaged in investigating the foundations of a science, we must set up a system of axioms which contains an exact and complete description of the relations subsisting between the elementary ideas of that science. [...] But above all I wish to designate the following as the most important among the numerous questions which can be asked with regard to the axioms: To prove that they are not contradictory, that is, that a definite number of logical steps based upon them can never lead to contradictory results. In geometry, the proof of the compatibility of the axioms can be effected by constructing a suitable field of numbers, such that analogous relations between the numbers of this field correspond to the geometrical axioms. [...] On the other

hand a direct method is needed for the proof of the compatibility of the arithmetical axioms.’ (Hilbert 1902, 437–79)

The question has interesting unfoldings for artificial intelligence developers. If it can be proved that the axioms of arithmetic will never be contradictory, by the use of these very axioms the system itself is self-provable, or complete. If a system can self-regulate itself without external, pre-defined premises, it obtains a kind of autonomy. But for computer science, being able to self-formulate the rules of a system remains a paradox. After all, software is nothing but a set of instructions. I could write a program that writes its own rules, but this would mean simply to create an overlapping set of rules that will define those. Would computer scientists be able to get away from this recursive loop, this closed chain of command?

At least for mathematics, the answer has been negative. This is the result of the investigation made by Kurt Gödel in his *Incompleteness Theorems*, originally published in German in 1931:

‘The most comprehensive formal systems yet set up are, on the one hand, the system of Principia Mathematica (PM) and, on the other, the axiom system for set theory of Zermelo-Fraenkel (later extended by J. v. Neumann). These two systems are so extensive that all methods of proof used in mathematics today have been formalized in them, i.e. reduced to a few axioms and rules of inference. It may therefore be surmised that these axioms and rules of inference are also sufficient to decide *all* mathematical questions which can in any way at all be expressed formally in the systems concerned. It is shown below that this is not the case, and that in both the systems mentioned there are in fact relatively simple problems in the theory of ordinary whole numbers which cannot be decided from the axioms.’ (Gödel 1992, 174)

Therefore, mathematics cannot prove itself, and will require some external rules — axioms — in order to build the scaffolding it will be based on. And if computers are contraptions that are based on the same rules of abstract logic as mathematics, this requirement can be extrapolated to state that computers will always need some external rule — a program, and with it, a *programmer*.

At this point in the history of abstract symbols, automated machine maths was still a pipe dream. But it is exactly this provocation by Gödel which would jumpstart the development of computers, as a young mathematician imagined a method to disprove another problem posed by David Hilbert.

Towards algorithmic art

The early twentieth century was not only prolific in new mathematical concepts. Artists were also pushing the limits of accepted support, proposing new expressive configurations. When Duchamp sent a urinal to a show in New York in 1917, he moved the focus of artistic skills from piece-making to concept-making. To use his terms, art should speak less to the ‘retinal’ and more to the mind (Judovitz and Duchamp 2010). To use a computing paradigm, conceptual art is closer to software than to hardware. It happens in the realm of ideas, where the concept is formed, to only later become a physicality that reflects it.

New steps towards algorithmic art were then proposed by participants of the Dadaist movement. In *To Make a Dada Poem*, Tristan Tzara offers instructions in the structure of a programmatic procedure:

Take a newspaper.

Take some scissors.

Choose from this paper an article of the length you want to make your poem.

Cut out the article.

Next carefully cut out each of the words that make up this article and put them all in a bag.

Shake gently.

Next take out each cutting one after the other.

Copy conscientiously in the order in which they left the bag.

The poem will resemble you.

And there you are – an infinitely original author of charming sensibility, even though unappreciated by the vulgar herd (Tzara 1920, sec. VIII).

Man Ray's piece *Object to be Destroyed* (1923), a metronome with a cut-out photograph of an eye on its balancing arm, also contained detailed instructions on the back:

Cut out the eye from a photograph of one who has been loved but is seen no more. Attach the eye to the pendulum of a metronome and regulate the weight to suit the tempo desired. Keep going to the limit of endurance. With a hammer well-aimed, try to destroy the whole at (sic) a single blow (Dreher 2005, 2).

Without the instructions, the piece would be incomplete, as the destruction proposed in the title would remain as a loose provocation. In 1957, a replica made by Ray on display in Paris was in fact destroyed by a crowd of protesting poetry students (The Theater: Battle of the Nihilists 1957).

Such examples cannot be considered the first to represent artistic expressions as procedural instructions. Hero of Alexandria's contraptions necessarily needed some structure that stored their progression of movements. The first forms of music notation date from ancient Mesopotamia, and they are also a representation of how to play notes in a sequence, deriving from the impossibility of registering the actual execution of songs until the late nineteenth century. But Tzara's instructions incorporate an element of randomness that was not consciously present in these previous cases.

Turing and the machine for mathematics

Hilbert continued to work on the most fundamental questions of mathematics for many decades. In fact, Gödel's solution was not written in response to the original twenty-three problems, but to the first question of a summary made by Hilbert at the International Congress of Mathematicians in Bologna in 1928:

- '(a) Is mathematics complete, in the sense that every mathematical statement could be either proved or disproved?
- (b) Is mathematics consistent, in the sense that a statement such as $2 + 2 = 5$ could never be arrived by (sic) valid proof, or in the sense that two contradictory propositions $a = b$ and $a \neq b$

b could both be derived?

(c) Is mathematics decidable, in the sense that there exists a definite method that can be followed to demonstrate that a mathematical statement is true or not?' (Dasgupta 2014, 45)

The last provocation is related to the tenth problem of the list compiled in 1900. It asked if there was a method by which to know if a type of equations known as 'Diophantine' had solutions in rational integers (i.e. zero, positive, and negative integers). This issue became known as the 'decision problem' (or *Entscheidungsproblem*). It was first resolved by Alonzo Church — another giant that laid the grounds of the science of computation — in a paper published in 1935 (Church 1936).

Church's solution hinted at programming methods. The paper describes lambda functions, which still play a fundamental role in programming today's computers. And it did raise the attention of Alan Turing, who had finished his undergraduate studies at King's College Cambridge with a thesis that qualified him to become a King's fellow. In early 1936, shortly after Church's paper, Turing published his own solution in the paper titled *On Computable Numbers, with an Application to the Entscheidungsproblem* (Turing 1937).

This is a seminal work because the method Turing devised uses what he called a 'Universal Computing Machine', now known as a Turing machine (Copeland et al. 2017). After centuries of research towards a representation system for reasoning, we finally get to a proposal that can be automated. And as we will see, the technology to implement that was being developed at the same time.

Turing imagined a machine that was composed of an infinite *tape* on which symbols could be written and read from. The symbols encoded values, and also the operations on the values. This is one of the most interesting effects of the exercise: to realise that operations could also be encoded as numbers on the tape, and decoded through the use of a *table of instructions*. The capacity to store a set of instructions is analogue to the storage of a program in the memory of present-day computers. The *head* of the machine represents the square of the tape that is being accessed at a given time, and a *state recorder* stores the current internal state of the machine, like initialising, finishing or reading. Today,

a system that is able to perform these tasks is known as a Turing-complete machine.

In very brief terms, the argument against the possibility of verifying that a solution to any given problem is possible or not resided in the fact that the machine would have to test all possible inputs. But since all inputs are stored as a sequence of numbers, like the programs, at a given point the machine will test the same program that is used to test all other programs, entering an inescapable loop.

Turing actually went out to pursue his PhD under the supervision of Church, at Princeton University, and returned to Cambridge in 1938. He was already interested in building a machine that would perform the theoretical exercise he imagined: some of the time he spent in the United States was used to build a binary multiplier machine (Hodges 2014). As the Second World War started, he joined the British cryptographic effort to break the cipher of the axis forces at Bletchley Park, the headquarters of communication intelligence in the United Kingdom. There, he devised several code-breaking methods that helped the Allied forces win the war.

He was also able to work there with actual computing machines. However, none of them had the flexibility he planned: they were built with specific pre-wired programs meant to decipher codes. Turing knew that electronics was the technology needed to run computers at a speed that would make the invention practical. Nevertheless, the machines built prior to the early 1940s were based on relays, a component that still had mechanical parts. Only in 1944 did he learn about the *Colossus*, the behemoth installed at Bletchley Park (Copeland et al. 2017).

After the war ended, Turing was able to focus on working on a *Universal Programming Computer*. He proposed a design focused on speed, named *ACE*. However, bureaucracy and politics caused his project to be delayed, while another computer named *Baby*, developed by former colleagues from Bletchley Park, was already functional at Manchester University in 1948. That prompted him to move there and collaborate on the design.

As with most great inventions, computers were being developed at the same time in different parts of the world. It is impossible to say which of them would be the first, as different criteria could be used to say what could be considered a computing machine as

known today. In 1938, electromechanical computers based on relays had already appeared. Konrad Zuse, working in Berlin with funds from Nazi Germany, developed a Turing-complete machine using this technology in 1942. IBM started to operate its Harvard Mark I in 1944, making calculations for the development of the atom bomb. The ENIAC, built by the US Army, was completely electronic and started functioning in 1945, although it was only after modifications made in 1948 that it could run stored programs. More machines were also developed in that decade, marking the beginning of the computer era as Turing had imagined. Even though they were slow and took the space of several rooms, the dream of a machine capable of synthetic mathematics was complete (Goldstine 1993).

Alan Turing was a remarkable character and is considered the inventor of computers as we know them today. He also made important insights regarding the concept of artificial intelligence, as we will later see. His life was tragically cut short by an apparent suicide in 1954, after being prosecuted and convicted on homosexuality charges.

Machines that think

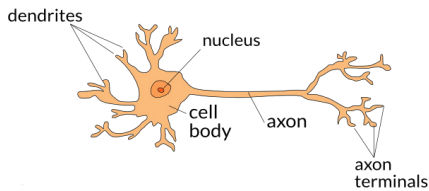
Finally, machines that could manipulate mathematical symbols were available, together with a representation system for them. And the speculation starts immediately: can machines think the way humans do? The 1940s were not only a prolific time for computer hardware: it was also the decade when the foundations for artificial intelligence were laid down.

In 1943, while Turing was working on breaking German cryptography, a ground-breaking paper was published. It dwelled exactly on the workings of the brain and proposed a mathematical model for the neuron that is still the basis for even the most current developments in AI, such as machine learning (McCulloch and Pitts 1943). The proposal was informed by the much earlier investigations of Camillo Golgi, which used his *black reaction* technique to yield photographic imaging of biological neurons, a representation with a clear symbolic similarity to the mathematical model. By using McCulloch and Pitts' equations, the biological neurons could be simulated by a machine, a theory that was still hypothetical at the time. In fact,

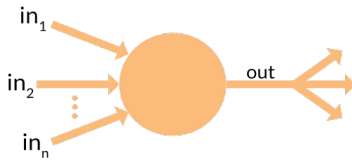
while the emulation of neurons has become commonplace with today's computers, the whole brain remains too complex for any device to imitate.

The model is of striking simplicity, and hence easy to implement: it just says that a given neuron will trigger an output if the weighted sum of its inputs surpasses a certain threshold.

Biological Neuron



Artificial Neuron



Adapted from Cristopher Berger,
<https://appliedgo.net/perceptron/>,
CC-BY-NC-SA

In the biological realm, this means that the axon terminal will raise its voltage once the sum of voltages of dendrites reaches a certain level. And in the computational simulation, this is modelled with numbers as input and output values. In current applications it would not be correct to brand the computation process as a simulation: on one side, neural networks generate very real outputs, and on the other side, besides this original inspiration, the models we will study in this thesis bear very little resemblance to the biological brain. But even if AI nowadays lies far from biology, the model of the neuron was a fundamental inspiration. As we will see, the field became split between two strategies. The first believed that the best way forward for AI was to manipulate logic symbols (snippets of information, like the analysis of the movement of a chess piece and all its possible consequences). The other believed knowledge is better grasped by training a neural network (feeding it the complete games with outcomes, for instance). The explosion of AI since the 2010s is evidence of the dominance of the latter strategy.

Cybernetics and systems theory

The 1940s brought not only the development of the first electronic symbolic manipulation machines; they also saw the birth of two frameworks of thought very important to this research: *cybernetics*, the science of artificial control systems, and *systems theory*, which seeks to understand the boundaries and exchanges between a system and its environment. By doing this, systems scientists face the difficult problem of defining the concept of autonomy of systems, which can be anything from living cells to computer programs.

A 1943 paper named *Behavior, Purpose and Teleology* outlined what would be baptised as cybernetics a few years later. It was written by Arturo Rosenblueth, Norbert Wiener and Julian Bigelow and discussed the concept of teleology in animals and machines, including speculation on how a robot would work (Rosenblueth 1943). Ideas present in the paper would still resonate decades later. When the paradigm of cybernetics shifted to systems theory, the reflection this paper proposed on the goals of objects or machines was still present: one way of defining the limits of a system is to think of its purpose. But the first outlines of systems theory were being drafted almost at the same time. In 1945 the biologist Ludwig von Bertalanffy published an article in German in which the term was first coined (von Bertalanffy 1945).

More texts of interest to this research appeared in this decade. Vanevar Bush published speculation on how machines could cooperate with human work in an essay that became a reference for computation, named 'As We May Think' (Bush 1945). And John von Neumann, who would become instrumental in the refining of computers as invented by Turing, practically founded Games Theory with the paper 'Theory of Games and Economic Behavior' (von Neumann 1944). This theory is still one of the best resources for creating platforms that allow machines to learn efficiently.

Von Neumann also discovered *cellular automata*, which are finite-state simulations of systems that are able to self-replicate and display persistent autonomous behaviours (von Neumann 2017). He proposed them while working on the problem of self-replicating systems, in particular robots that could build perpetuating replicas of themselves. Stanislaw Ulam, his colleague at the Los Alamos National Laboratory, suggested that these structures should be

discrete in order to be able to design a reductionist model of life (Ilachinski 2002). The study of cellular automata became a field of its own, and will be revisited later on in the text, when I describe John Conway's *Game of Life* in the context of the artistic component *Crank of Life*.

Founding cybernetics

Starting in 1946, a series of conferences were organised in New York City and became known as the Macy Conferences, after the foundation that hosted them. The meetings were proposed as an interdisciplinary gathering, where medicine and biology could be mixed with mathematics and concepts such as *information, teleology and self-regulation*, or even how computers could learn (ASC: Foundations: History of Cybernetics n.d.). In fact, the title of the first meeting in March of that year was 'Feedback Mechanisms and Circular Causal Systems in Biological and Social Systems', and it was attended by, among others, Margaret Mead, Gregory Bateson, Julian Bigelow, Warren S. McCulloch (chair of the conference), John von Neumann, Norbert Wiener and Arturo Rosenblueth.

The Macy Conferences were seminal in setting the scene for the emergence of artificial intelligence as a research area, even if the term would only appear much later. Ten such meetings were organised in total, with subjects like information theory, homeostasis, decision theory and complexity. Influenced by the conversations held in the conferences, Wiener wrote a book that followed up on his 1943 behaviour paper, proposing the concept of cybernetics, or the science of self-governance. In 'Cybernetics: Or Control and Communication in the Animal and the Machine', Wiener (1948) looks at the autonomy of biological beings and wonders how it can be emulated by a man-made device. One of his focuses is on the mechanisms of self-control, namely negative feedback loops: systems that try to keep a dynamic system within certain bounds, like the thermostat of an oven, or the buoyant valve of a toilet flush. He also covers a variety of subjects that would be of great relevance to AI and computer science, such as the binary number system (which still was not a standard), information and its relations to bandwidth and noise, and even the possibility that machines could play chess.

In a second edition published 13 years later, he added a few chapters that expanded the discussion on the self-reproduction of machines and the self-emergence of patterns in evolutionary processes, as well as learning processes within the brain, aspects that are central to the discussion of self-organisation. The author also addresses issues that are still very current, such as the risk of job losses in the face of automation.

With this work, Wiener not only founded cybernetics as a science; he also made it popular, as most parts of his book were very accessible to laypeople. He wrote a book a few years later with the intention of reaching an even broader audience, named 'The Human Use of Human Beings: Cybernetics and Society'. The Macy Conferences themselves did not escape the feedback effect and were renamed, from 1950, as *Cybernetics Conferences*, after a suggestion from the then chair Heinz von Foerster (ASC: Foundations: History of Cybernetics n.d.). On the other side of the Atlantic, in a similar but more informal setting, cybernetic minds gathered to dine in London at the Ratio Club from 1949 until 1958. Founded by John Bates, it had Turing, W. Ross Ashby and Donald McKay among its attendees (Husbands and Holland 2008).

Besides Wiener's classic book, 1948 also saw the publication of Claude Shannon's paper on information theory, *A Mathematical Theory of Communication* (Shannon 1948). This work laid down some basic concepts in machine communication such as emitter, channel and receiver. Even the most basic unit of information used in computer science, the *bit*, is named there for the first time. It represents a binary state — true or false, 1 or 0, yes or no. The combination of 8 bits form a byte, which can store numbers from 0 to 255 by the combination of binary states. Shannon also brought together concepts of thermodynamics and cybernetics by proposing the concept of *information entropy*, which represents the level of uncertainty in a message; after a certain level of entropy, the original message cannot be decoded anymore. The proportion of *noise* in relation to the *signal* carried by the channel must be kept low. If this is not possible, the system should use *redundancy*, repeating essential symbols so that loss can be afforded without missing the message. Shannon detached the concept of meaning from the message; all that matters in his theory is how much information the channel can carry and

still be decoded at the other end. In his sense, nonsense content is just as valuable as a news report: ‘Frequently the messages have meaning; [...] These semantic aspects of communication are irrelevant to the engineering problem.’

Cyberneticists applied many of their findings to the field of organisation management: Stafford Beer wrote *Cybernetics and Management* (Beer 1959). He was hired by the Chilean government of Salvador Allende to develop the Cybersyn project, a national computerised economic management system. These ideas were recently reclaimed by Web3 designers of decentralised autonomous organisations (DAOs) as anarchist tools (Zargham et al. 2023).

The beginnings of systems theory

Many of the ideas around the organisation of systems emerged in the late 1940s and early 50s. In 1947, psychologist W. Ross Ashby published *Principles of the Self-Organizing Dynamic System* (Ashby 1947), or systems that are able to reorganise themselves in response to changing environmental conditions. In 1950, Bertalanffy published his *Outline of General Systems Theory* in English (von Bertalanffy 1950), practically founding the science of systems.

‘The formal correspondence of general principles, irrespective of the kind of relations or forces between the components, leads to the conception of a ‘General Systems Theory’ as a new scientific doctrine, concerned with the principles which apply to systems in general’ (von Bertalanffy 1950).

Systems theory aims to explain how different parts of a system interact and affect each other, as well as the system as a whole. As such, it has been applied in fields as diverse as engineering, psychology, biology and sociology. It can be used to analyse complex structures such as a power grid or an aeroplane, as well as biological organisms, cells, economic and social systems. This attempt at being universal also makes a closed delimitation of the theory elusive to date. ‘Systems theory is a term without a formally agreed upon definition’ (Adams, Hester, and Bradley 2013). Again, it is helpful for this research in as much as defining systems also presupposes defining how autonomous they are, that is, where they stand in the continuum between self-sufficiency and dependency from other systems.

An approach to sociology from the systems perspective was published by Talcott Parsons as early as 1951 in the book *The Social System*. In 1954, a first meeting of Bertalanffy and other scientists invested in the field took place in Palo Alto. This group would found, in 1956, the Society for General Systems Research (renamed to International Society for Systems Science in 1988).

This focus on self-control and later self-organisation also intertwined with the field of cybernetics. In fact, it was usually difficult to separate the fields since the scientists working on these issues could at times be associated alternatively with one, the other or even both sciences. But as cybernetics evolved and became more associated with artificial intelligence, systems theory continued to stand as a field of its own. Bertalanffy stressed that cybernetics should be seen as a particular subfield of the latter, more than a competing worldview:

‘Systems theory is frequently identified with cybernetics and control theory. This again is incorrect. Cybernetics as the theory of control mechanisms in technology and nature is founded on the concepts of information and feedback, but as part of a general theory of systems [...] [t]he model is of wide application but should not be identified with “systems theory” in general [...] [the] warning is necessary against its incautious expansion to fields for which its concepts are not made’ (von Bertalanffy 2009, 17-23).

The imitation game or Turing test

Although not directly associated with either camp, Turing was already wondering how the machines being developed could compare to human standards of intelligence. His 1950 paper ‘Computing Machinery and Intelligence’ (Turing 2009) starts by proposing the question ‘Can machines think?’ and goes on to propose an ‘imitation game’. His concern is to provide a platform where a machine can be interrogated without giving away its non-human condition: ‘No engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin’. Therefore, the questions and answers in the game are typewritten, and the participants are

sitting in isolated rooms. Turing's example is to have a man (A), a woman (B) and an interrogator (C). By asking questions, C must find out who is the man and who is the woman.

“We now ask the question, ‘What will happen when a machine takes the part of A in this game?’ Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, ‘Can machines think?’ Turing 2009, 434).

This game is now known as the *Turing test*. Its use as a measure of machine intelligence is controversial to say the least; every day there are more programs that pass the test with flying colors and still bear no signs of human-like wit, from the Elisa ‘psychiatrist program’ in the 1970s to the recent conversations with Google’s LaMDA chatbot (Heng-Tze Cheng 2022). But Turing himself saw it more as a philosophical than a computing problem, related to Ada Lovelace’s questioning of whether machines can originate anything. One of its most famous rebuttals is the *Chinese Room* (Searle 1980), a thought experiment proposed by philosopher John Searle. As in Turing’s test, the subject being questioned is isolated in a room, receiving questions and passing answers through typewritten papers. It must convince its interrogator that it is not a machine but a person who can communicate in Chinese. Searle asks us to suppose that a given program manages to pass the test and to suppose that the machine is replaced by a person who cannot speak Chinese, but has the complete instructions to answer the questions in English, and thus can replicate its workings. This person will also pass the test, yet has no understanding of Chinese or of what it is being asked about. Therefore, Searle’s question is whether such a computer is actually intelligent or is it only a simulation of intelligence, and proposed a difference between ‘Strong AI’ (the former) and ‘Weak AI’ (the latter). ‘... according to Strong AI, the correct simulation really is a mind. According to Weak AI, the correct simulation is a model of the mind’.

Artificial intelligence appears

By the early 1950s, war efforts were long over. After being used for military projects (the first programs run by the Harvard Mark I and ENIAC were nuclear weapon simulations), several computers were available for research on both sides of the Atlantic. Chess and draughts programs were written for Manchester University's Ferranti machines (Copeland 2020). One interesting project was the 'Logic Theorist', (Crevier 1993) developed in 1955. It was a program capable of manipulating mathematical expressions and solving problems, which was able to solve 38 of the first 52 theorems in the aforementioned 'Principia Mathematica', tested on a machine belonging to the RAND Corporation. It was the dream of a machine capable of symbolic reasoning coming true, even if it was very limited.

At this point, the term *artificial intelligence* had not yet been coined. It is interesting that when it finally appeared, it happened in the context of a reaction against the neural approach praised by the participants of the Macy Conferences, and also as an alternative science to cybernetics itself. It happened at a workshop hosted by Dartmouth College in the summer of 1956, organised by John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon. The term appeared for the first time in the proposal sent out to request funding for the meeting. The creators of 'Logic Theorist', Allen Newell and Herbert Simon, attended the workshop and presented their invention (McCorduck 2004), which was warmly welcomed, perhaps exactly because it did not use any neural techniques, but a kind of brute-force technique: 'reasoning as search'. This consists of exploring and testing all possible solutions to a problem and has very little to do with the intuitive human methods of problem-solving. It is also notable that Marvin Minsky, who later came to be a staunch critic of neural methods, had previously developed one of the first devices that implemented a neural network, the SNARC (Crevier 1993). Another attendee, Arthur Samuel, became known for releasing, in the same year, the first draughts game that was reasonably challenging, or 'tricky but beatable' by his own account (Mitchell 2019; Samuel 2000). Samuel was also the protagonist of an argument related to the aforementioned *Lovelace Objection* with Norman Wiener, when the latter stated that machines

‘could develop unforeseen strategies’ in game programs (Wiener 1960). Samuel sides with Ada with the rebuttal ‘[...] the machine will not and cannot do any of these things until it has been instructed as to how to proceed’ (Samuel 1960).

The term picked to downplay neural techniques is now used mostly to designate these very strategies: as we will see, since the 2010s artificial neural networks have dominated the field, and symbolic techniques are now the exception in artificial intelligence research.

Symbolism, subsymbolism

This is a good moment to clarify some of the terms used in this thesis. There are several approaches to solving the question of how to create reasoning machines. *Genetic programming*, for instance, is a field that has yielded interesting results. Others have developed ideas around *artificial life*, which is very close to systems theory and is inspired by natural life processes. Nevertheless, the dominant methods in the history of AI have been the one described before: one based on the manipulation of symbols and the other based on simulated neuron networks. It is hard to say which came first. As we have seen, there is a centuries-old tradition of trying to build a symbolic representation of thought, while the first mathematical model of neurons appeared only in 1943. But the rift was probably first made visible by the organisation of the Dartmouth workshop described above.

The line of thought based on symbolic manipulation became known as symbolism, and was the most common strategy in the years thereafter, to the point that it is also called GOFAI – good old-fashioned artificial intelligence (Haugeland 1986). Every definition would be an oversimplification, but for the sake of this thesis I will assume it uses rule-based inference, logical agents, and does not rely much on statistics (although it does incorporate probabilistic methods as a way of avoiding failure) (Russell and Norvig 2021). The idea underlying the symbolic framework is that all human thinking can be replicated by logical processes, being replaced by ‘symbols, combinations of symbols, and rules and operations on symbols’ (Mitchell 2019). After all, the investment in logical systems bore

some cases of success, like the Deep Blue chess computer that beat the human world champion, Kasparov, in 1997. But by the mid-1980s, expert systems — symbolic AI approaches that rely on humans to create rules that reflect expert knowledge of a particular domain — were increasingly revealing themselves to be brittle: that is, error-prone and often unable to generalise or adapt when presented with new situations (Boden 2018). In this project, symbolic AI may also be referred to as GOFAI.

Subsymbolism, on the other hand, is based on methods that use neural networks, borrowing inspiration from the structure of the brain. From the 1970s on, the term connectionism also began to be used, referring ‘to the idea that knowledge in these networks resides in weighted connections between units’ (Boden 2018). Yuk Hui describes connectionism elegantly:

‘This effort to go beyond formal representation of perception and reasoning is known as connectionism, represented by a movement in cognitive science that uses artificial neural networks to explain cognitive capacities, and it resonates with the twentieth-century philosophical attempt to go beyond representation’ (Hui 2021, 238).

The term originated in psychological research, and is closely related to learning processes. The field of connectionism also led to what is known today as machine learning. The demise of symbolism and the dominance of subsymbolism happened in part because the processing power needed for such processes became recently available. I will refer to this faction as connectionism, subsymbolic AI or machine learning.

Symbolic/subsymbolic schools are also related to System 1 (fast) and System 2 (slow) types of thinking, proposed by psychologist Daniel Kahneman (Kahneman 2011). System 1, or fast thinking, is the unconscious, instinctive way of processing information. It happens for instance when trying to locate the source of a sound, read a slogan or process simple sentences. Slow thinking happens when we need to focus and give our attention to a task, like solving an equation, counting how many instances of the letter A are in a text, or parallel parking into a small space. So when our brain forms

System 2 thoughts, humans are manipulating symbols using logic — the car controls, the letter A, mathematical variables — much like symbolic AI aims to solve problems. And subsymbolic AI processes information a bit like how fast thinking works: by learning from exposure to raw data, without reasoned explanation of how we know what we know. We don't need to make a logical evaluation to determine where a sound comes from or if an object is coming in our direction: we arrive at the conclusion quickly and without any conscious reflection.

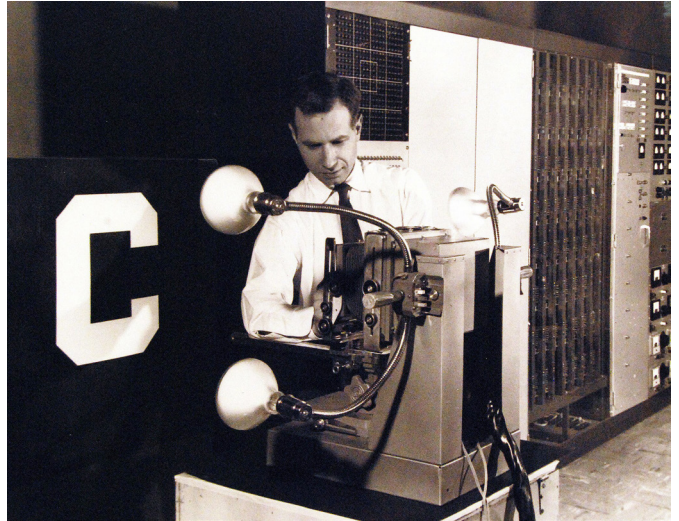
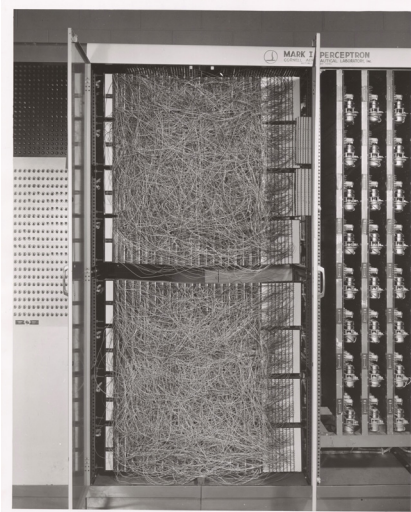
An important footnote is to mention that all neural AI done today is done on computers, and therefore requires programming, which is nothing other than symbol manipulation. So in some way, symbolism is unavoidable in the subsymbolic methods of today. However, symbolic programming is there mostly just to put the neural networks to function; the actual artificial intelligence work is done subsymbolically.

Perceptrons

One of the best-known breakthroughs in AI happened on the connectionist side in 1958: the invention of the perceptron machine by Frank Rosenblatt. Even if the concept of the perceptron had emerged in the McCulloch and Pitts paper in 1943, no successful implementation of an artificial neural network had been seen until then. The model was first tested on an IBM 704 computer, but like many of the computer science projects of the time, Rosenblatt's device was a single-purpose machine, designed to be trained to recognise visual shapes. It consisted of an array of 400 photocells, creating a rudimentary 20×20-pixel camera. These were connected to potentiometers — an analogue electrical device that is used to dim lights or control the output volume of a radio, for instance. The setting of each potentiometer was used to physically store the weights of the artificial neurons. During the training process, electric motors would adjust their positions.

The success of the perceptron machine derives from its ability to recognise some basic shapes, an absolute novelty for its time. Its failure also consists of the incapacity to recognise most shapes. The flaws overshadowed the achievements: a few decades after its

inception, the history of AI became synonymous with symbolic IA. Reclaiming subsymbolism would start in the 1980s, with some success stories in the 1990s, until a complete dominance of the field in the second decade of the 21st century.



By unknown photographer, presumably working for Cornell Aeronautical Laboratory. Original publication: Cornell University Library, Fair use National Museum of the U.S. Navy

It is interesting to see how much, at that moment, the research turned away not only from the neural path but also from learning processes, focusing on symbolic representation. A paper written in 1958 by McCarthy defends the development of a knowledge representation that can be manipulated easily (McCarthy 1960). In Minsky's own words:

'The most central idea of the pre-1962 period was that of finding heuristic devices to control the breadth of a trial-and-error search. A close second preoccupation was with finding effective techniques for learning. In the post-1962 era the concern became less with "learning" and more with the problem of representation of knowledge (however acquired) and with the related problem of breaking through the formality and narrowness of the older

systems. The problem of heuristic search efficiency remains as an underlying constraint, but it is no longer the problem one thinks about, for we are now immersed in more sophisticated subproblems, e.g., the representation and modification of plans' (Minsky 1968, p. 9, cited in Buchanan 2005).

But the coup de grâce against neural methods came in a book by Minsky and Seymour Papert published in 1969, named *Perceptrons*. The book criticises Rosenblatt's research, and was 'interpreted as showing that further progress in neural nets was not possible, and that this approach to AI had to be abandoned' (Olazaran 1996) due to the fact that perceptrons cannot calculate the XOR function described earlier in this chapter. That is not exactly true. Rosenblatt had already shown how these calculations could be done using different configurations (Rosenblatt 1961). The book itself mentions how McCulloch and Pitts proposed to create a Turing machine capable of all Boolean operations. What Papert and Minsky proved is that a specific architecture of artificial neuron networks could not be used for these operations; a much more complex set of perceptrons would have to be used, with several layers. They arrived at the same conclusion when investigating the issues of parity (the ability to identify whether the number of connected neurons is even or odd) and connectedness (the ability to distinguish between a figure and the background in images): the complexity of the networks required for these tasks was too large for the existing technology at the time the book was written (Olazaran 1996). In any case, Minsky has expressed scepticism about the potential of 'Gamba networks' (the name he used for multi-layer perceptrons) on other occasions, and the book did cause the funding for this approach to go dry in the 1970s (Alom et al. 2018). The dismissal of connectionism and the difficulties inherent to GOFAI led to what is known as the first AI winter (Buchanan 2005). It is possible that the perceptron controversy delayed the development of machine learning. But it is also a fact that the deep learning revolution was contingent on the development of computers with capable processing power.

Art, cybernetics and systems

As cybernetic ideas became popular, an increasing number of artists began incorporating them into their work (Shanken 2002a). British cyberneticist Gordon Pask combined scientific production with creative inventions. From 1953 until 1957, he designed and displayed *Musicolour*, a piece used in theatre plays that translated sounds captured by a microphone into coloured lights (Dreher 2020, 38). Another pioneer was Hungarian-born artist Nicolas Schöffer, who designed kinetic sculptures that reacted to their environments. His work *CYSP1* (1956) used phototubes and microphones as sensors that triggered movements of the piece itself (mounted on motorised wheels) and its parts. Both Schöffer and Jacques Bureau, the designer of the circuits that controlled the movements of *CYSP1*, use W. Ross Ashby's term *homeostasis* to describe its ability to 'move and adapt to external events' (Dreher 2020, 44). Other works of his expanded these concepts towards the public space, building interactive towers reaching tens of metres, like *Chronos 8* (1968) (Gehér 2018). However, the autonomous traits of *CYSP1* point towards the cybernetic desire for an artificial organism, and in this case, one that engenders a particular choreography.

In the 1960s, cybernetic art took the shape of a large-scale artistic movement. Roy Ascott (United Kingdom, 1934), one of its exponents, fully adopted a cybernetic worldview for his work.

'The art of our time tends towards the development of a cybernetic vision, in which feedback, dialogue and involvement in some creative interplay at deep levels of experience are paramount' (Ascott 1968, 105–12).

Ascott's works were not so concerned with the autonomy of the artwork, but they incorporated concepts of loop and feedback (Shanken 2002a). His piece *Change Paintings* (1959), a set of paintings on plexiglass, allowed viewers to manipulate and reconfigure parts of the painting.

In his thesis, Lynch proposes a distinction between *cybernetic* and *systems* artists of the time (2018, 102). Following this categorisation, Hans Haacke (Germany, 1936) would belong to the latter group. He explored systems in nature, beyond the human-made cybernetic

sense. The piece *Condensation Cube* (1965) consisted of an isolated glass cube with moist air inside. By increasing and decreasing its temperature, water would evaporate and condensate, inciting scientific investigations into systems thermodynamics, input-output exchanges and preservation of energy. Although trained as an artist, Jack Burnham (USA, 1931), became more influential as an art historian and theorist. He developed the concept of system aesthetics (his original orthography) as a method to integrate system studies into art practice (Burnham 1968).

The exhibition *Cybernetic Serendipity* (1968) captured the infatuation of artists in the field, displaying works that related computers to art. Curated by Jasia Reichardt (Poland, 1933) after a suggestion by Professor Max Bense (Reichardt 1969, 5), it opened at London's Institute For Contemporary Arts and later travelled to different cities in the United States (Shanken 2002a). Gordon Pask's *Colloquy of Mobiles* offered torches as a way for visitors to interact with its moving parts; Nam June Paik (Korea, 1932) presented the remote-controlled robot K-465, in collaboration with engineer Shuya Abe (Japan, 1932). Sculptures guided by light and sound were contributed by John Billingsley and Edward Ihnatowicz, respectively.

In 1970 Burnham explored his own concepts of system art in the exhibition *Software, Information Technology: Its New Meaning for Art*, which opened at the Jewish Museum in New York City. It focused more on the relationship between the metaphysical aspect of conceptual art and software: "The show drew parallels between the ephemeral programs and protocols of computer software and the increasingly "dematerialized" forms of experimental art, which the critic interpreted, metaphorically, as functioning like information processing systems' (Shanken 2002b). Whether Burnham viewed this relationship as metaphorical is up for debate: both conceptual art and software exist only as abstractions that affect the audience in various sensuous ways (Burnham 1974, 47).

Also relevant for this thesis is the fact that at this moment, artists began to explore graphics created with the aid of computers. This implies the elaboration of algorithms that define images, which again brings software close to conceptual art. Visual artist Sol Lewitt (USA, 1928) proposed for his series *Wall Drawings* to be built according to written instructions, like: 'Black arcs using height

of the wall as a radius, and black arcs using the midpoints of the wall as a radius. The arcs are filled in solid and drawn in India ink' or '500 vertical black lines, 500 horizontal yellow lines, 500 diagonal right red lines, 500 left blue lines within a 72' square'. The idea of software-defined art is used by Philip Galanter for his concept of generative art (Galanter 2016), while LeWitt himself compares the method to the notations of score music (Roberts 2012).

In 1962, Michael Noll, working at Bell Laboratories in Murray Hill, NJ, USA, was shown a plotted graph gone astray by a colleague that resembled a concrete abstract drawing. Noll, who in tandem with his engineering education had a strong interest in art, decided to investigate the use of the plotter to purposely create abstract paintings. He describes the atmosphere of the laboratory as intellectually stimulating, with other engineers also investigating the use of computers for artistic production. After writing and testing different plotting programs, he found one result particularly pleasant, where the positions of lines were randomly defined (Noll 1994). He named it *Gaussian Quadratic*, and, with the support of the lab, tried to register the copyright for the work with the Library of Congress Copyright Office. As we will see again later in the research, US copyright law does not posit computer-created works as copyrightable, and the registration was rejected. After two more attempts he managed to get it registered, but he had to waive the initial claim that the work was a collaboration between man and machine, and state the work was done completely by him (Zeilinger 2021b, 61).

At the time, two theories of computer aesthetics were being developed concurrently in Europe. While Abraham Moles was writing his *Informational Theory and Esthetic Perception* (Moles 1966), Max Bense developed his computational aesthetics (Bense 1998). Even though Bense was primarily concerned with algorithms for aesthetic evaluation, some of his students like Geg Nees (Nees 1969) and Frieder Nake (Nake 1968) started creating procedural graphics with the help of computers and plotters.

Around the same period, Vera Molnar co-founded the group *Art et Informatique* at the Parisian Institut d'Esthétique et des Science de l'Art, and started her own algorithmic explorations (Molnar 1975). Herbert W. Franke published his first computer-aided visual works in 1967, although he had been experimenting with mechanical

analogue methods for drawing and photography since the 1950s (Franco 2022).

The history of generative art is vast, with a production that closely reflects the developments of computer and telecommunications technology. It is instrumental in developing *autopoiesis* within art. A definition of generative art that supports this claim comes from Galanter: ‘Generative art refers to any art practice where the artist uses a system [...] which is set into motion with some degree of autonomy contributing to or resulting in a completed work of art’ (Galanter 2003).

However, for the purpose of this thesis and in the interests of objectivity, I will only return to it when the use of artificial intelligence tools enters the scene, in the 2010s.

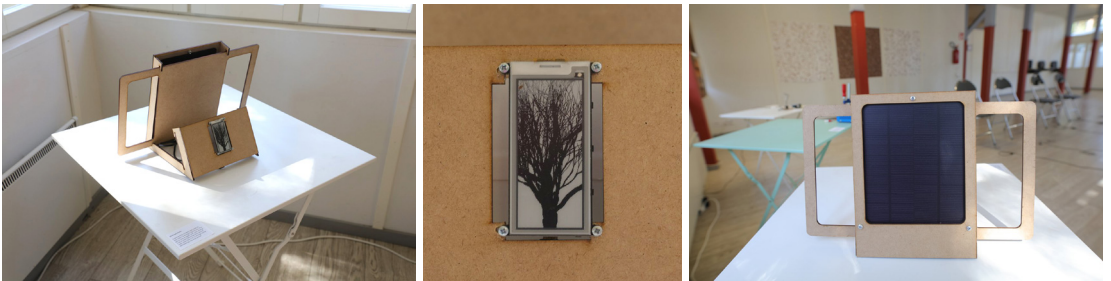
Artistic component: Photosynthesis

The piece *Photosynthesis* does not incorporate neural networks in its system. Within the development of AI, it would fit among the symbolic efforts to reproduce human cognition. Paradoxically, it might be the most autonomous of the works in the project, capable of permanently drawing unique sketches of trees without human intervention, as long as it is lit by the sun’s rays. This also demonstrates the fact that artificial intelligence platforms require much more energy spending than traditional computing, making them less autonomous in terms of their operation in the proportion that its perceived intelligence increases. Another artistic component, *Aiku*, which will be described later, will substantiate the limits of circumstantial autonomy.

The piece consists of a support holding on one side a small solar panel, and on the other a black and white e-paper display, which is attached to a single board computer, model ESP32. When the sun hits the panel, it produces enough electricity to start the computer, which is programmed to draw tree-like structures on the display. E-papers have the double advantage of consuming very little energy and preserving the last drawn image, even when it has no power. They are common in e-readers such as the Amazon Kindle, and they’re not suited for moving pictures, as the update speed is too slow. In *Photosynthesis*, each tree takes about 30 seconds to be produced and exhibited.

Providing its own energy, and drawing according to its own rules, this piece brings my work closer to the concept of autopoiesis than any other. *Photosynthesis* lives in its own enclosure and establishes distinct relations with the environment, by absorbing the sun's rays and outputting artworks that in their turn have different effects on the gallery visitors. Granted, its aesthetics did not emerge but were designed by another entity — myself, the artist/observer outside the system's limits.

The code to program the ESP32 device is written in C, with a driver library to plot on the e-paper. The algorithm is based on a CC-licensed open processing sketch by Asher Salomon. The code is published under the same licence.



Photosynthesis (2021). Photos by Tanja Becher, used with permission.

Second-order cybernetics and autopoiesis

The year 1967 marks an inflection in cybernetics research. In a keynote speech at the meeting of the American Society for Cybernetics, Margaret Mead called upon its members to apply cybernetic thinking to cybernetics itself, inaugurating what came to be known as second-order cybernetics (von Foerster 1969, chap. 1). This leads to a rethinking of purposes within the field. This renewal was strongly advocated by researchers who worked on ethics from a cybernetics perspective of self-reflection, like von Foerster (von Foerster 2003).

This inflection spawned newly renovated theories that brought new ways of understanding the cybernetic paradigm. Gordon Pask proposed his *conversation theory* (Pask 1976), while in general the development of sociological systems-based thinking was strongly disseminated. Niklas Luhman is known for applying a systemic approach to sociology, law and the arts (Luhmann 1995, 2000;

Luhmann, Ziegert, and Kastner 2004). Gregory Bateson, after a life-long contribution to anthropology, takes a holistic approach to ecology and society through the gaze of cybernetics (Bateson 2000).

Of particular interest to us is the appearance of the Santiago School of Cognition, around the findings of Humberto Maturana and Francisco Varela. In a 1972 book, they propose the concept of *autopoiesis*, a cybernetic theory of life that helps us delimit what it means to be autonomous.

This concept describes a self-referential network of processes that continuously produces or maintains the components that make up a system. Different from a car factory, which produces a system (the car) other than itself, a cell or a school of fish is able to support its homeostatic condition by a series of intricate relations with its surroundings. Living things are able to cope and adapt to the changes in their environment, sustaining their individuality as distinct and coherent entities.

These boundaries between the organism and the outside are fluid, yet well-defined. The exchanges through this boundary allow the thing to keep its operational closure; that is, the system is defined by its own capacity for autonomy.

Maturana and Varela also make the case that all autopoietic systems bear some kind of cognition; this cognition is defined by its capacity to understand itself within the environment and find the materials needed to self-support. Using a strictly biological definition, these would be nutrients, sunlight, oxygen. But when applied to a social construct, the cognition is defined by the elements that build themselves and also build their context — culture, references, language and affections as a result. Several art scholars have adopted the concept of autopoiesis as a referential framework for art (Bishop and Al-Rifaie 2016; Bishop and al-Rifaie 2017; Sevänen 2001), even if Maturana explicitly stated that the concept only applies to biological sciences.

Artistic component: Crank of Life

In 1970, a column published by the Scientific American magazine made public a particular *cellular automaton* that brought together

various concepts in systems, cybernetics and computation. *Game of Life* was invented by British mathematician John Conway. ‘Because of its analogies with the rise, fall, and alternations of a society of living organisms, it belongs to a growing class of what are called “simulation games”, games that resemble real-life processes’ (Gardner 1970). The article gave instructions on how to play the game on a paper board, but it was the availability of computers later that made *Game of Life* so popular and so studied.

It is a game that plays by itself. All that has to be done is an initial configuration, which consists of markers on a grid. These markers represent the initial entities that are ‘alive’. After that, the following rules must be evaluated in every round:

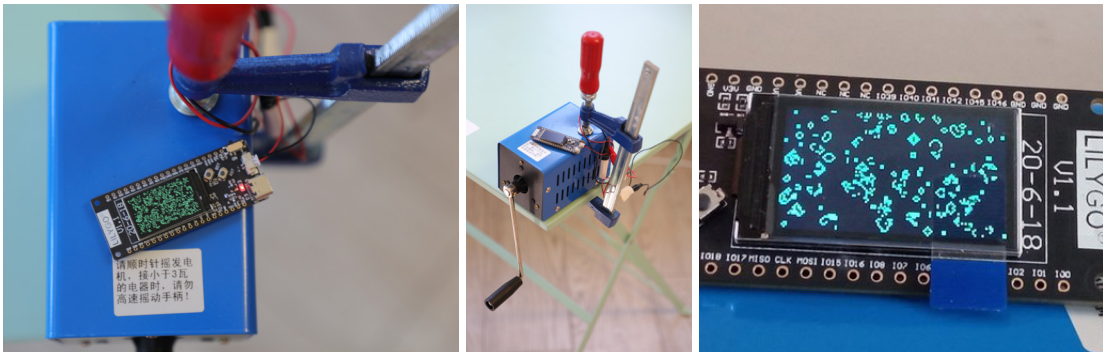
- “1. Survivals. Every marker with two or three neighbouring counters survives for the next generation.
2. Deaths. Each marker with four or more neighbours dies (is removed) from overpopulation. Every marker with one neighbour or none dies from isolation.
3. Births. Each empty cell adjacent to exactly three neighbours – no more, no fewer – is a birth cell. A marker is placed on it during the next move.”

(Gardner 1970)

These basic rules allow very complex patterns to appear. Some repeat themselves in a loop, some move around the board and might leave different patterns in their wake. The affordance of complex shapes and behaviours from strict rules is considered one of the best examples of emergence and self-organisation (Bak, Chen and Creutz 1989). When patterns touch, interesting interactions occur, with specific outcomes. One specific pattern, which was named ‘glider’, displays persistent shapes as it moves across the screen. It allowed Conway to show how logic gates such as AND, OR, NOT and memory counters could be implemented in the game, making it Turing-complete: within it, any kind of computation can be performed (Berlekamp, Conway and Guy 1985).

Conway’s invention is the subject of my artwork *Crank of Life*, presented as part of the Research Pavilion exhibition in 2021 at the Hietsun Paviljonki in Helsinki. The piece proposes an outer

enclosure that adds a new layer to the game system. The game is stored in a program for the LilyGo programmable chip (which is also based on the ESP32 platform) that has a small colour LCD screen. As in the case of *Photosynthesis*, the system is not fed by an external electrical source. In order for the simulation to start, one must use a crank that generates electricity through a dynamo. When the chip is booted, a random cell configuration is drawn on the screen. The randomness is given by the voltage provided by the dynamo at the moment the chip starts running. Therefore it is determined by the viewer, as they can apply more or less weight to the lever.



Crank of Life (2021). Photos by Tanja Becher, used with permission.

The piece discusses the limitations of agency on so-called autonomous systems. The choice of colour and position of the display are reminiscent of biological cultures in Petri dishes. By conditioning the start of the simulation to a God-like human movement, we stress how non-emergent artificial life and intelligence are: play things, contingent on supplied energy, intelligent design, and made for our own intellectual delectation.

The code to program the ESP32 device is written in C, with a driver library to draw on the screen. The algorithm is based on an implementation of the Game of Life made by Joan Soler-Adillon for the Processing platform. The code is published under a Creative Commons licence.

AI winter, developments of subsymbolism and feats of symbolism

After the falling into disrepute of perceptrons, there came the period known as AI winter. Throughout the 1960s, funding for AI investigations came in the order of millions of dollars each year for many academic institutions such as MIT and Carnegie Tech. But progress was slow and results were few. For instance, the Speech Understanding Project by the research agency DARPA was shut down after five years and fifteen million dollars for failing to achieve useful speech recognition (Crevier 1993).

Several reasons contributed to this downfall. Problems of scale plagued the implementation of symbolic AI systems: the brute force strategy of searching all possible solutions, used by the ‘Logic Theorist’ and its successor, the General Problem Solver (GPS), came to a limit when the ‘combinatorial explosion’ made the number of possibilities so vast that it could not be handled by the computers of the time. It is very hard to create a representation for common sense knowledge differently from mathematical theorems. This led to what is called the Moravec paradox, first formulated in 1988: ‘It is comparatively easy to make computers exhibit adult-level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility’ (Moravec 1995). A more succinct phrasing of the problem is proposed by Steve Pinker: ‘The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard’ (Pinker 2010). In other words, a very complicated calculation can be done in a fraction of a second by a computer. Yet an easy task for a human, such as recognising a face or a voice, is very difficult for a program — or at least it was until recently.

The rebound would happen around the late 1970s and early 80s, fuelled by the success of expert or knowledge systems. These were platforms focused on solving specific problems, like MYCIN, which diagnosed bacterial infections, or XCON, to configure VAX model computers, sold by DEC (Digital Equipment Corporation) (Kautz 2022; Russell and Norvig 2021). New optimism took over the research, as well as the funders. A project was initiated to tackle the issue of non-specific, common-sense knowledge — the expertise

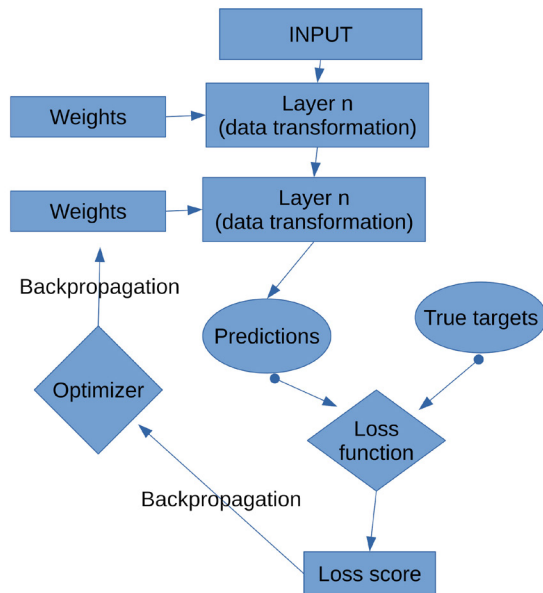
disposed of by every human, or the ‘easy things’, like knowing that rivers run towards the sea, or that every mammal has a mother. Cyc, as it is named, was started by researcher Douglas Lenat in 1984, and it still exists as a company today. Its database contains more than 1,500,000 entities, 24,500,000 relations between them, and it is permanently being updated (Lenat, Prakash, and Shepherd 1986).

After a decade without significant research due to the *Perceptrons* book backlash, subsymbolism also made a comeback in the 1980s with some notable advances. A very important work that reclaimed the potential for the approach were the two volumes of *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, by David Rumelhart and James McClelland (1986). It provided a clear framework for systems based on neural networks, focusing on its parallel processing aspect, but setting in stone important features such as the activation function and output function. But probably the most significant concept to be developed at that time was *backpropagation*.

The idea of backpropagation is closely related to *feedback*, the concept so dear to the cybernetics pioneers. But it is a fundamental piece of artificial neural networks which had not been established until that moment. Some predecessors had indeed appeared before, in the 1960s and 70s, especially for control applications (Goodfellow, Bengio and Courville 2016). Seppo Linnainmaa is considered to have done the first implementation of a backpropagation algorithm in 1970 (Griewank 2012). But, as with most great inventions, it was developed almost simultaneously in different institutions around the mid-1980s. In 1982, Paul Werbos applied Linnainmaa’s method to a neural network (Schmidhuber 2015). The results appearing in the aforementioned Rumelhart book came from a collaboration with Geoffrey Hinton (Rumelhart, Hinton and Williams 1986), and Yan LeCun also used the algorithm in his thesis project (LeCun 1985). LeCun and Hilton shared the 2018 Turing Award for computer science for the work that started with these advances, together with Yoshua Bengio.

So how does backpropagation work? Neural networks must be trained to execute a given function, like recognising images of dogs. This example would use a set of images known to be of dogs, and another set of non-dog images. The network functions by feeding

the images to the input neurons and watching the output neuron: an output value close to 1, for instance, indicates that the network is confident it has been fed a dog image, as opposed to a value near zero. In the training, what the backpropagation does is adjust the weights (a numerical value given to each connection between neurons) when the network guesses wrong, either a false positive or a false negative. The evaluation of this assessment is done through the loss function, which calculates a score depending on how well the results correspond to the correct output. An equation named gradient descent indicates how the weights must be adjusted, until the system learns to recognise dog images with the greatest percentage of success possible. This is still the core of today's machine learning systems.



Training with backpropagation

Another important piece of the deep learning machinery was also being developed around the same time. It was again inspired from research in the biology of the nervous system, like the perceptron. Neurophysiologists David Hubel and Torsten Wiesel started a collaboration in the late 1950s to understand the processing of information

in the visual cortex. They found that mammalian visual perception works in layers. The first neurons to receive information are triggered by shapes, such as edges. As the information is feed-forwarded further into the brain, more complex objects and scenes make the latter neurons activate more often (Hubel and Wiesel 1959, 1962). Their work and other investigations on visual perception led them to be awarded the Nobel Prize in Medicine in 1981. They also inspired a Japanese engineer, Kunihiko Fukushima, to develop what he called the Cognitron (Fukushima 1975), an artificial neural network organised in layers, and its successor, the Neocognitron (Fukushima 1980), which had some success at recognising handwritten digits.

The kernel of Fukushima's idea is to use a convolution to transform the data between one layer and the other. This is a matrix operation that works in two ways: it highlights the specific features that a given layer is interested in, and it reduces the dimensions of the image.

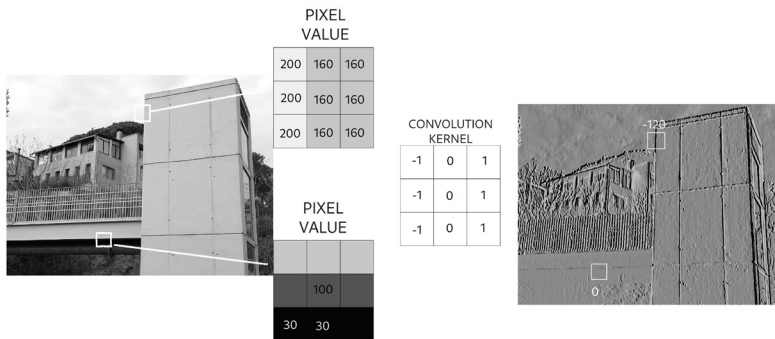
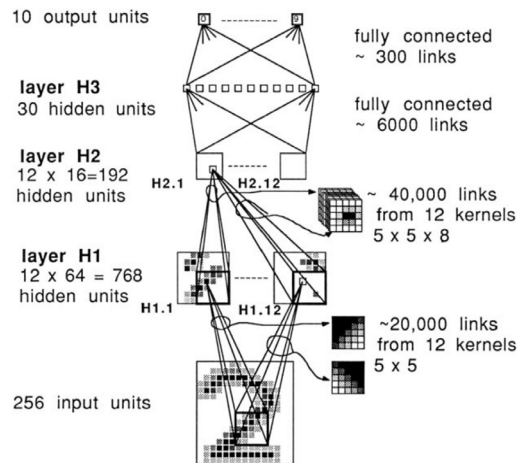


Illustration of how convolutions are used to detect vertical edges. For example, a convolution of the upper receptive field with the weights is $(200 \times -1) + (160 \times 0) + (160 \times 1) + (200 \times -1) + (160 \times 0) + (160 \times 1) + (200 \times -1) + (160 \times 0) + (160 \times -1) = -120$. The other area will yield a result of 0, as there is no vertical edge.

In the example in the figure, the convolution highlights the areas with vertical edges between clear pixels on the left and darker pixels on the right. By combining a great number of such units, the convolutional network is able to identify the most common features in the images it is looking for. After the calculation, the nine original numbers in the grid will become just one in the next layer. This reduction will happen in several of the layers in the network.

The networks using these methods were named ConvNets or CNNs (convolutional neural networks). The combination of back-propagation and CNNs were used by Wei Shang in 1988 to recognise alphabet letters, and by Yan LeCun in 1989, in an application to read handwritten postcodes (LeCun et al. 1989; W. Zhang et al. 1988). LeCun's software was deployed in commercial use from 1996 (LeCun et al. 1998).

It is worth taking a moment to look at LeCun's 1988 network, not to be confused with the 1998 network which he baptised LeNet. It is one of the most basic configurations of NNs, and it is used widely as a teaching example, and consists of an input layer, an output layer, and three layers in between them. The neural layers other than input and output are always called hidden layers. Today, a network that consists of only five layers is considered very simple, but it required a considerable amount of computing power at the time.



From (LeCun et al. 1989)

This figure should be read from the bottom, where the input image is represented (a handwritten number 2), to the top layer consisting of ten output elements, one for each digit from 0 to 9. The input image size is 16×16 pixels, in greyscale, with a total of 256 pixels. The figure shows that a convolution operation happens between the input layer and layer H1: it uses the patches

of 5×5 pixels (the kernels) of the input to calculate the pixels in the H1 layer. Two different convolution weights are shown for feature maps H1.1 and H1.12, represented by the two squares with 5×5 dimensions. The maps on H1 are only 8×8 pixels wide, demonstrating the compression that happens as we move forward through the layers. These ‘pixels’ on the other layers are referred to as units. The picture depicts the number of links between the layers. It is not worth getting into the details of the connections here; what is important for the network is that the values of some units in a layer are combined with others through a calculation and used to determine the value of the unit it is connected to. This calculation is the sum of the multiplications of the output value from the previous layer by its weight — and the weight is exactly what gets updated through the backpropagation method. In this network, a bias value is also added to the sum. The weights and bias are initialised with random values, which will be adjusted by the backpropagation/training process.

The twelve feature maps in layer H2 are only 4×4 pixels large. These are fully connected to the 30 units in H3, that is, all 192 units in H2 are connected to all 30 units in H3. H3, in its turn, is fully connected to the 10 output units that will bear the values indicating the probability of which the digits from 0 to 9 were represented in the input image.

The history of neural networks is not complete without the history of the data used to train and test the models. The selection of data may convey biased choices and cause misleading results. Around this time, the first public datasets began to appear, like TIMIT, a speech collection which gathered American voices with different accents. LeCun trained this experiment with handwritten numbers collected by contractors hired by the United States Postal Services.

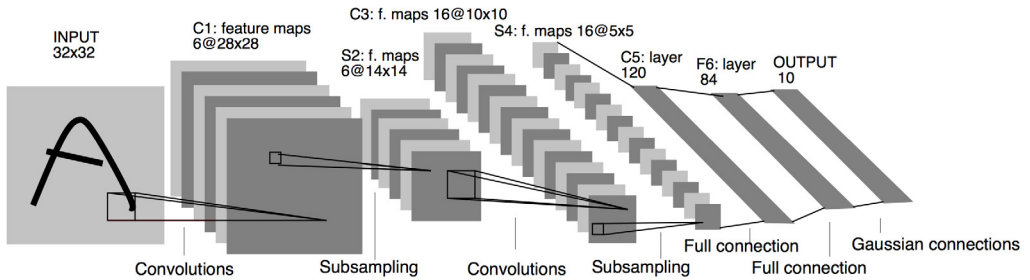
The goal of LeCun’s experiment was to prove that backpropagation could be used to train neural networks, which was accomplished by achieving a success rate of 95%, meaning that it only identifies incorrectly 5% of the test set. Despite the progress within connectionist approaches, there were few practical or commercial applications of artificial intelligence. Expert systems had reached their limit: they were hard to update (since most were based on special

hardware, at a time when personal computers started to be widely available), ‘brittle’ (subject to gross mistakes when input data is outside the expertise of the system) and could not learn new skills (Crevier 1993). A second AI winter arrived, from the late 1980s to the early 1990s (Strickland 2021). What rescued the field was the rapid evolution of computers: with faster machines, more processing power could be thrown at the problems.

This is how a phenomenal win for AI was achieved. In 1997, for the first time, a man-made chess-playing contraption defeated the human world champion in a series of six games. This was actually a rematch. The first series of games, in 1996, had been lost by Deep Blue by four games to two. The second time, it won three games and drew one. The system started to be developed at Carnegie Mellon University in 1989, and from its release in 1996 to the winning version, the main difference was a two-fold increase in speed (Campbell 1999; Newborn 2003).

Deep Blue’s approach to chess playing was a classic symbolic, GOFAI one. It was essentially not different from the Logic Theorist, making an extensive search on a tree of possible moves. Granted, it had extensive knowledge of chess strategies and more than 700,000 games played by grandmasters. Its database had, for instance, all the endgames with five pieces or fewer, and several six-piece endgames. It didn’t exactly exhaust all possibilities, since it uses heuristics to determine which of the paths are not worth exploring. But it does rely on having the computing power to do this massive exploration of possible outcomes.

Connectionism also continued its steady development. A new version of LeCun’s number classification network was released, this time named *LeNet-5*. It had 345,308 connections between neurons and 60,000 trainable parameters. The network error rate was only 0.7%. The technique had been in use by the company NCR in many banks across the United States since 1996, and the documentation was published in a 1998 paper. The paper was also the debut of an historical dataset of handwritten numbers named MNIST (Modified NIST), which the authors built by combining two existing databases from the National Institute of Standards and Technology (LeCun et al. 1998). This set is still widely used for teaching and benchmarks in the machine learning community.



Architecture of LeNet-5, from (LeCun et al. 1998)

Describing the history of artificial intelligence after 1998 is more difficult, not because of another winter or crisis, but simply because it became commonplace. The domestic vacuum cleaning robot Roomba was released in 2002. DARPA's grand challenge for autonomous vehicles started in 2004 and by 2005 it already had several cars driving autonomously for more than 200 kilometres. The second generation of Furby toys, released in 2005, had a speech recognition system that allowed them to have basic conversations.

The banalisation of artificial intelligence is also related to the concept of *AI Effect* (Haenlein and Kaplan 2019). It seems like once a big AI problem is solved, there comes a perception that it is not an AI problem anymore — it is just plain computation. Another synthesis is provided by Larry Tesler: 'Intelligence is whatever machines haven't done yet' (Hofstadter 2000). Once a computer beats the best human chess player, it becomes transparent that it is just brute force applied to the problem, and that the program doesn't think like humans do. Granted, this is very true — human cognition and the machine way of solving problems and accomplishing tasks are not comparable, no matter whether it is a neural network or a GOFAI system. We don't examine millions of possible outcomes before deciding on a chess move; and the way images are created from text by *Stable Diffusion* models is completely different from the process of an artist fulfilling a commission, let alone coming up with a new style. The difficulty comes from our own definition of intelligence, and perhaps it highlights the difficulties caused by the term AI itself,

which forces a permanent comparison with biological intelligence — which, in its turn, no one knows exactly what it is. ‘Looked at in one way, everybody knows what intelligence is; looked at in another way, no one does’ (Sternberg 2000). The concerns raised by Searle’s Chinese Room appear again. Defining intelligence as a mere attribute of a system is misleading, as it is a quality distributed across domains — the system, previous art, the body. At the same time, designing a system that appears to bear specific intelligent qualities is not difficult, and it has been done since the times of the original Mechanical Turk.

GENERATIVE VISUAL AI

In the 2000s, artificial intelligence development seemed to have stalled. Neither symbolic nor subsymbolic fields were bringing new advances to the tasks yet to be solved by computers. As we have seen, one of the most important problems that hindered automation in several industries was the difficulty of making systems that understood visual information. However, creating a machine that could check for defects in manufactured products or to read an address on an envelope seemed to be out of reach.

The state of the art of machine vision in the 2000s was the field known as computer vision. The guidebook for the most common vision platform of the time, OpenCV, only had one final chapter on machine learning — the remaining techniques described are all symbolic procedures (Bradski and Kaehler 2011). The main impulse for the neural field at the time came from a competition named ImageNet, created around the dataset that goes by the same name.

ImageNet

ImageNet takes its inspiration from WordNet, which is a dataset of words from the English language organised in categories, and the semantic relationships between them (Fellbaum 2010). Its visual counterpart was started at the initiative of Professor Fei-Fei Li at Princeton University, who realised that a collection of categorised images could be helpful in developing image classification algorithms (Gershgorn 2017). WordNet was conceived as a tool to train symbolic artificial intelligence into semantics, but ended up facilitating the subsymbolic turn of the 2010s by originating ImageNet.

Li's project was to download millions of images from the internet and have them manually categorised into a database. The project started in 2007, but only took off when she learnt of Amazon's Mechanical Turk, a distributed remote work platform. Other options were explored, such as hiring undergraduate students for the job, but they turned out to be very expensive or wouldn't scale. Using

the Mechanical Turk allowed them to distribute the task to thousands of workers, and doing it quickly (and at an affordable price, at the cost of cheap labour). However, many difficulties remained: the categorisations had to be double-checked and verified by different algorithms, as humans may also make labelling mistakes. In 2009 ImageNet was finally released, but it received little attention until the idea for a challenge came along (Deng et al. 2009).

The complete set was first launched with 3.2 million images in 5,247 categories, but the challenge used a subset with 1,247,167 images in 1,000 categories. Some 50,000 were used on top of these as validation, since the training of neural networks requires references, and another 100,000 were used as a test set for the challenge itself, to measure the efficacy of each competing solution. However, the original set continued to expand. In 2023, it had already gathered 14,197,122 images in 21,841 categories.

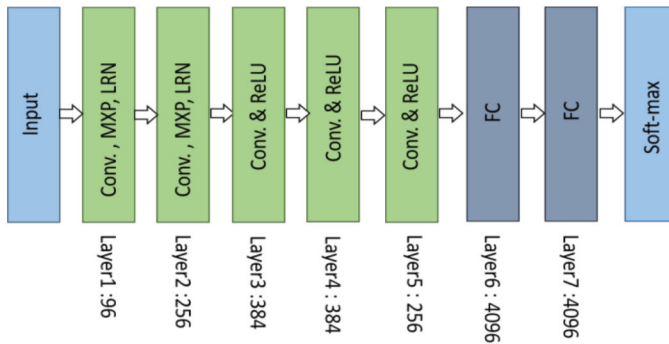
Assembling such a huge amount of real-world data also implies capturing the existing problems of society, unless specific counter-measures are put in place. In the case of ImageNet, these hitches lurked in the data for many years. In 2019, Kate Crawford and Trevor Paglen published *Excavating AI*¹, a report on the biases and troubling labels found on ImageNet and other sets. They found classifications with ‘problematic, illogic and cruel’ labels. Under the human female subtree there can be found, for instance, ‘gold digger’ or ‘ball-breaker’. A photo of Barack Obama labels him as ‘antisemite’, while another of him holding a Nazi flag is described as ‘Bolshevik’. These mishaps have been fixed, but others arose: an investigation by Vinay Prabhu found a large number of pornographic pictures (Prabhu and Birhane 2020), which were also subsequently removed. A later effort from the ImageNet team blurred all recognisable faces from the object set to protect the privacy of the subjects (Yang et al. 2022). As we will see, current models have started to use generative AI to create the visual training data, which brings another set of troubles of its own.

In computer science, open challenges and competitions are important tools for fostering new developments. They provide a standardised framework against which new solutions can be

¹ <https://excavating.ai/>, accessed 28 August 2023

measured and compared. In this case, the framework is the big set of images which must be categorised correctly by computational methods in relation to human-made labels. The first competition on ImageNet came out in 2010, when the winning team obtained an error rate of 28.2 per cent, still using a procedural method. The turning point would come in 2012, when a deep convolutional neural network, AlexNet, was first used in the challenge and obtained an error rate of only 16.5 per cent (Krizhevsky, Sutskever and Hinton 2017). This network has similarities with LeNet and other previous architectures, but vastly outstrips them in terms of scale, with a total of 60 million parameters.

This challenger was also one of the first to take advantage of the development of GPUs, a processing chip that until then was used mostly for 3D graphics. The type of matrix calculations needed for 3D rendering is very similar to the ones used in neural networks, therefore all that was needed was to write software that could run NNs on them. The popularisation of GPUs also paved the way for cryptocurrency experiments.



AlexNet architecture. from The History Began from AlexNet (Alom et al. 2018), CC-BY

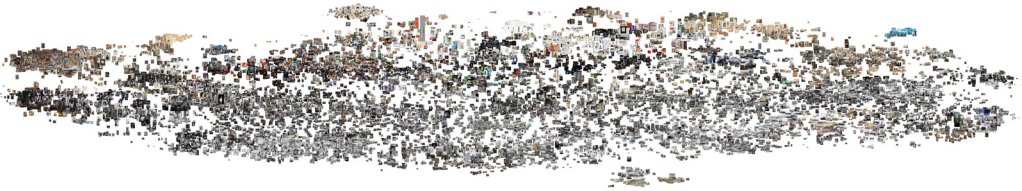
As early as in 2013, almost all teams were using a deep learning approach (Russakovsky et al. 2015). In 2015, ResNet-152 won the challenge with an error rate of only 3.57%, surpassing the human rate at 5% (Alom et al. 2018). After that the classification problem was considered effectively solved, with the last competition taking place in 2017, even before ImageNet's ethical issues surfaced.

Nevertheless, even though more varied datasets have surfaced like Laion-5B, ImageNet is considered an industry standard and continues to be used for visual machine learning. The will to explore this artificial intelligence landmark was my own motivation for the artistic piece *Supervision*.

Artistic Component: Supervision

In 2022, the exhibition *Paradoxes of Photography* opened at the Finnish Museum of Photography. It was funded by the *Olga and Vilho Linnamo Foundation*, and was part of the *Post-digital Epistemologies of the Photographic Image (PEPI)* research project, funded by the Academy of Finland. As such, I was invited to present a proposal for an artistic intervention that incorporated materials from the museum image archive.

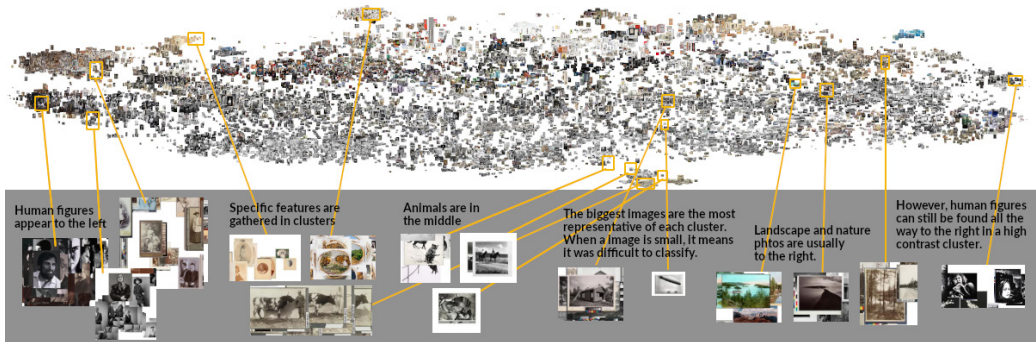
Armed with curiosity about the particular effect that the dataset ImageNet had (and is still having) on visual perception and generative culture, I set off to explore the archive with analytical tools based on this set. The result was the disposition of all the 12,684 pictures provided by the museum on a tableau with a width of six metres and a height of one metre. .



Supervision, 2022. Digital print, 600x100cm

The result has a cloud-like structure that reveals how ImageNet organises the archive in its categorisation, complete with semantic relationships between the represented objects. But zooming in on (or walking closer, in the case of the gallery) this organisation displayed a sense of the world that quite often uncovered biases contained in the database. The cluster of generic non-Western people,

for instance, discloses its colonial perspective. These clusters are also visually marked by the size of photos. The most representative example of each category is larger (after a suggestion from the curator) and stands in the middle. This allowed the appearance of blank spaces as separation between clusters.



A small guide with some selected *Supervision* clusters.

The sheer number of photos and the historical character of the collection allowed the audience to engage in games of exploration. Many examples have an affective importance to a Finnish audience, presenting known historical figures, public personalities, recognisable artworks and places.



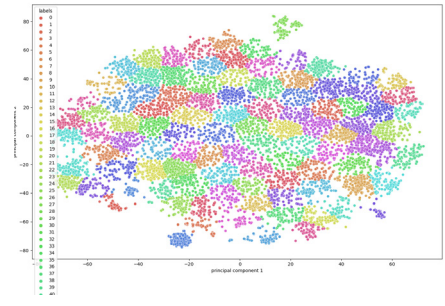
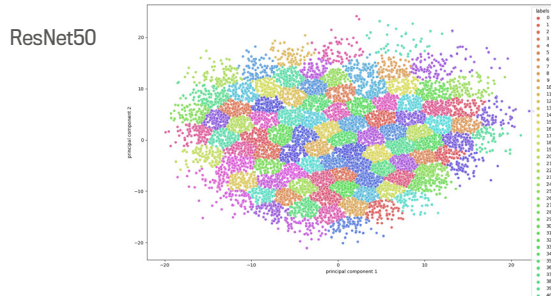
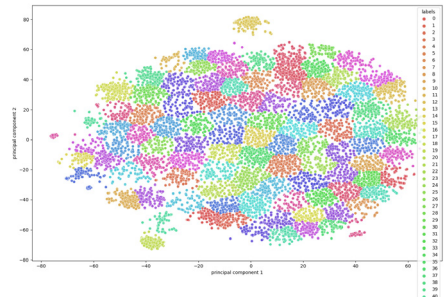
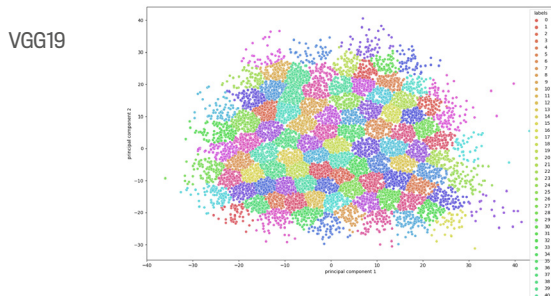
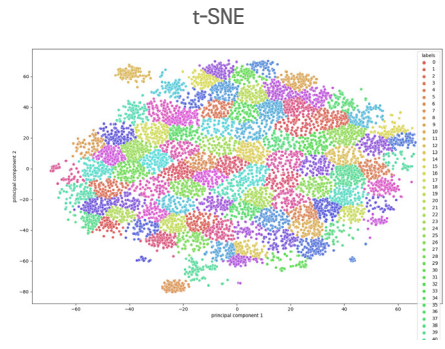
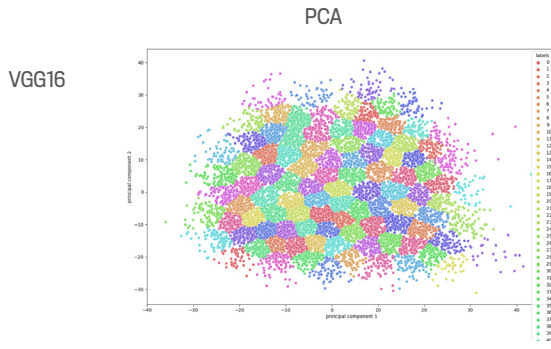
Close-ups of the tableau.

The process consisted of using a model trained on the ImageNet set to perform an operation on each image. The operation outputs a set of numbers, the embedding vector, which is unique for every image and represents their relation to the ImageNet categories. But in order for this representation to be displayed in a meaningful way, another operation must be performed on every embedding, namely a dimensionality reduction, since the original resulting vector has 2048 dimensions. Firstly, a direct reduction to the bi-dimensional x and y locations of a Cartesian plane, using Principal Component Analysis (PCA) was attempted, with poor results (Tipping and Bishop 1999). Then a two-step approach was followed, with PCA being used to reduce to 100 dimensions, and after that another reduction to two variables was performed using the t-SNE method (t-distributed stochastic neighbour embedding) (Hinton and Roweis 2002). After tweaking parameters to make full use of the horizontal space of the tableau, an algorithm named KMeans (Lloyd 2006) was applied to calculate the representativity of each image in its cluster, which was then visualised as its size.

Other visualisations were explored, like a fixed grid or the Voronoi algorithm, which divides areas in order to create a representation of the clusters of different categories (Voronoi 1908). Different models like VGG19 and VGG19 were also used (these yield 4096-dimensional embeddings) (Simonyan and Zisserman 2015). It is worth mentioning that the OpenClip model, which will be reviewed when we address text-to-image platforms, should provide a more natural, continuous segmentation of subjects, less based on separate classes. But the ResNet50 model offered the possibility to use weights pre-trained with ImageNet (He et al. 2015). For its historical relevance and the opportunity to study and reveal its shortcomings, this was the chosen approach. The code is published on the author's Github repository² and can be used by anyone willing to visually organise a large collection of pictures.

2 <https://github.com/brunovianna/collectionview/>, accessed 28 August 2023

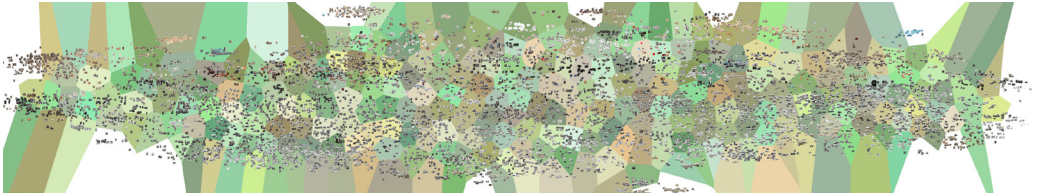
DIMENSIONALITY REDUCTION X MODEL
(full size graphics available on the Research Catalogue)



VISUALISATIONS



Grid



Voronoi

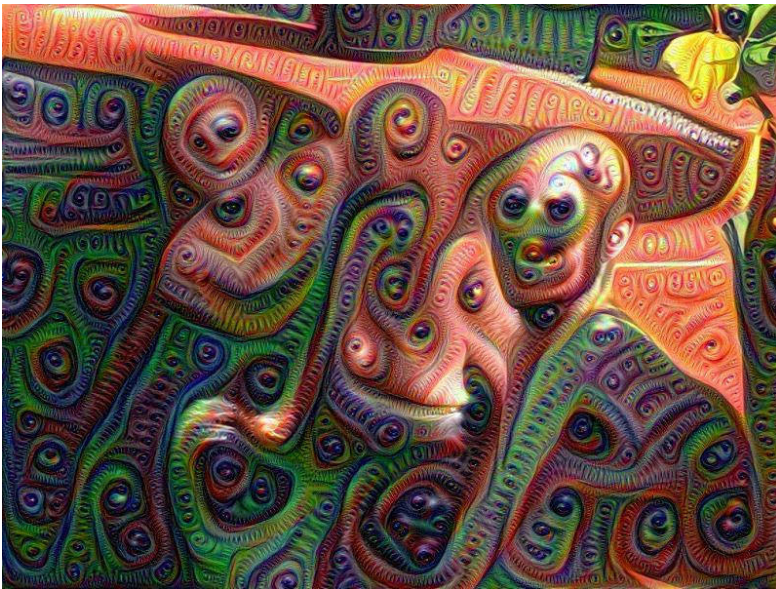


Final
(weighted sizes)

After a further commission from the museum, the project unfolded into a video-based version, where the image cloud floats in a 3D space, and different clusters are visited by a virtual camera. The piece became part of the museum's permanent collection.

DeepDreaming of Electric Sheep

The first visual instantiation of neural networks that drew attention for its aesthetic potential was DeepDream. It originates from a network developed by Google for the 2014 edition of the ImageNet challenge (Szegedy et al. 2014). It was first published in their research blog in 2015, and consisted of a convolutional neural network trained on specific images (such as dogs or places). When another image was fed into it, the outcome was what the researchers described as ‘beautiful art’³. The process was given the name of ‘inceptionism’ and became popular enough to lead to the first commercial tools, like deepdreamgenerator.com⁴ in the same month.



A deepdream image created by Jessica Mullen, CC-BY 2.0

Another neural experience that made the rounds was the encoding of the movies *Blade Runner* and *A Scanner Darkly* (Romano 2016).

3 <https://web.archive.org/web/20150708233542/http://googleresearch.blogspot.co.uk/2015/07/deepdream-code-example-for-visualizing.html/>, accessed 28 August 2023

4 <https://web.archive.org/web/20150719235130/https://deepdreamgenerator.com/>, accessed 28 August 2023

Terence Broad, a student at Goldsmiths, University of London, trained an autoencoder on the movies as his final thesis project for a master's degree in creative computing (Broad and Grierson 2016).

Autoencoders were invented in the 1980s and are a well-known idea in the field of deep learning (Goodfellow, Bengio and Courville 2016, 499). It is a network that tries to copy its input to its output, passing through a smaller hidden layer. This has the effect of compressing the original data, but in a way that it can later be decoded. A method published in 2015 used a GAN-like discriminator to train the network to obtain results most similar to the input as possible (Larsen et al. 2016). The movie reconstructed from the frames encoded in the network created using this method was published on YouTube, and it was similar enough that the artist received a copyright infringement notice. But the results draw attention mostly for the uncanny visuals, which allude to the original movies in a ghostly, eerie fashion.

Generative adversarial networks

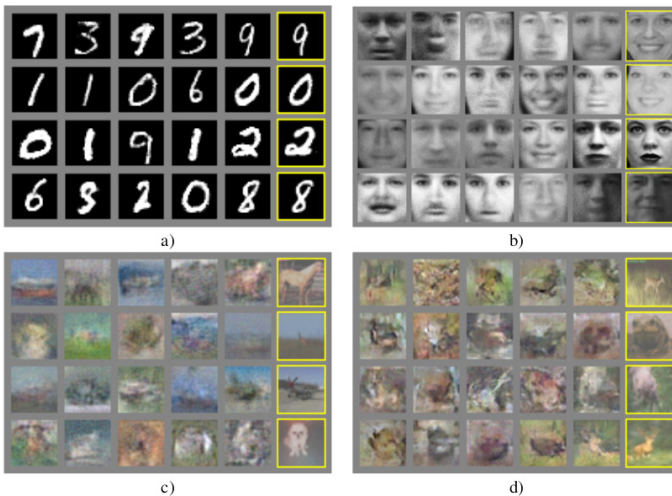
Generative art took a vertiginous leap with the invention of generative adversarial networks (GANs) in 2014: instead of generating what could be described procedurally, creators could spawn anything that could be 'taught' to a network if there was a large enough number of examples to be learnt. This tilts the scales one more step towards the autonomy of the artistic system, from the programmed randomness of procedural towards an exploratory, 'mapping the unknown' stance. Before GANs, generative art was limited to what could be expressive through a symbolic algorithm (Caldas Vianna 2020).

As the story goes, Ian Goodfellow, a doctoral student at Université de Montreal, was celebrating the graduation of another student in a bar when he was asked to help solve a computing problem. Their colleagues needed to fit a large amount of training data on a specific model of GPU, a graphical processing unit, which at the time could only fit 1.5 gigabytes of data. After a few beers, he had an idea and went home to try it (Giles 2018). Instead of fitting the massive network, he would combine two smaller ones, trained with the same data, competing against each as in a game. The first would be trying to generate images, while the other one would judge

their quality. The ratings produced by the latter would help with the training of the first one.

‘(The discriminator) estimates the probability that a sample came from the training data rather than (the generator). The training procedure for (the generator) is to maximize the probability of (the discriminator) making a mistake’ (Goodfellow et al. 2014).

As with many innovations, this solution was also imagined by another person at more or less the same time. Olli Niemitalo proposed it in a blog post in 2010, but never implemented it (Niemitalo 2010). Other approaches using dual networks have been used before, in 2006 and 2013 (Abu-Khalaf, Lewis, and Huang 2006; Wei Li, Gauci and Gross 2013). The fact is that Goodfellow produced recognisable images on the same night, which became immensely popular after the publication of the results in the relevant paper, which includes an acknowledgment of the bar where the algorithm was conceived (Goodfellow et al. 2014).



Visualisation of random results obtained using the GAN method (in yellow) compared to previous methods with different training sets - a) MNIST, b) TFD, c) CIFAR-10 (fully connected), d) CIFAR-10 (convolutional discriminator and 'deconvolutional' generator) (Goodfellow et al. 2014)

Since then, machine learning systems have started to adopt the multi-network approach in more and more complex configurations. Today, such systems are huge amalgamations where outputs from some networks become inputs for others, sometimes mixing up heterogeneous data modes like text, images and sounds.

GAN explosion

The publication of Goodfellow's code and paper triggered an explosion of tweaks, experiments and customisations to the original idea. Besides image generation, GANs were developed for different applications such as designing DNA, the generation of music and speech, discovering drugs and many others (Gui et al. 2020).

The evolution of image GANs is of particular interest for this work, as they inspired several artists that took upon the tool and developed different creative practices. These could include from an installation built around generated images to models specifically trained to produce a given style or subject. What is particular about GANs is that, when trained with a consistent set of pictures, they will spawn new images in the style in which it was trained. These new outputs will be in the same style but uniquely different from all images belonging to the training set.

The network built by the training process constructs a *latent space*, a collection of possible outputs, which is mapped to a vector of numbers. In the BigGAN model, for instance, this latent vector has 128 numbers ranging from -1 to 1 (Brock, Donahue and Simonyan 2019). Every little variation in even one of these numbers will yield a different image. Similar images are grouped in vectors with a short numerical distance between their components.

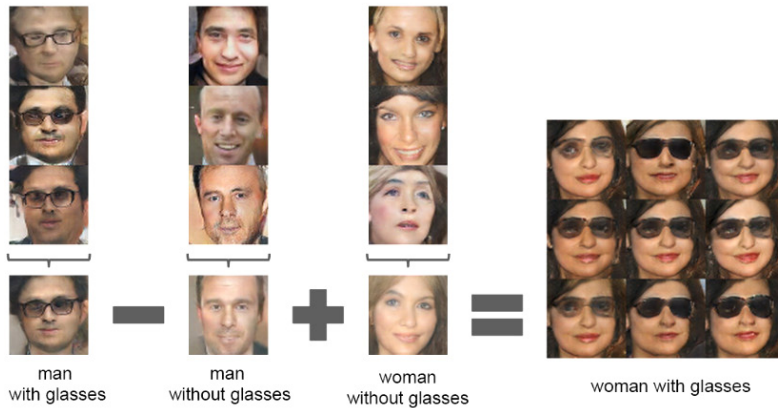
This allowed a wide range of experiments. Networks trained on faces have regions in the latent space where particular features can be found, like gender, ethnicity, facial expression and in-between combinations. Animations were made from one point in the latent space to another, creating various morphing effects. A network trained on photos of cats and dogs would not only spawn variations within the species but also all sorts of cat/dog hybrids. The last developments of GANs before the appearance of diffusion models were trained on generic sets, allowing them to be used in text-to-image systems.

The amount of GAN variations is huge. A repository⁵ dedicated to collecting GANs, last updated in 2018, listed 501 variants at the time. I will describe here only a few of those that allow for the creative generation of images, as they are more relevant for the project.

DCGAN

DCGANs are germane to this section, as we will see that they were the basis for some of the first art-based experiments on GANs. They were invented in 2015 and published in 2016. The improvement they offered was the ability to train at a large scale, obtaining better quality in bigger images (Radford, Metz and Chintala 2016). The repository⁶ by Soumith Chintala claimed the developers tried scaling up to 4096×496 pixels, but results larger than 256×256 do not look very good.

It also proposes the walk in latent space, as already described, and vector arithmetic. This technique consists of performing mathematical operations on the latent vector to obtain the corresponding effect on the images. For instance, ‘smiling man’ minus ‘neutral man’ plus ‘neutral woman’ results in a smiling woman.



Arithmetic operations on latent vectors (Radford, Metz and Chintala 2016)

5 <https://github.com/hindupuravinash/the-gan-zoo/>, accessed 28 August 2023

6 <https://github.com/soumith/dcgan.torch/>, accessed 28 August 2023

Pix2pix

This solution proposed the use of GANs to transfer features from one image to another. A satellite picture of a neighbourhood could be translated into a map, a daytime scene into night, black and white to colour. The publication of the method in the author's repository⁷ in 2016 allowed many artists to carry out experiments. The paper came out in 2017 (Isola et al. 2018).

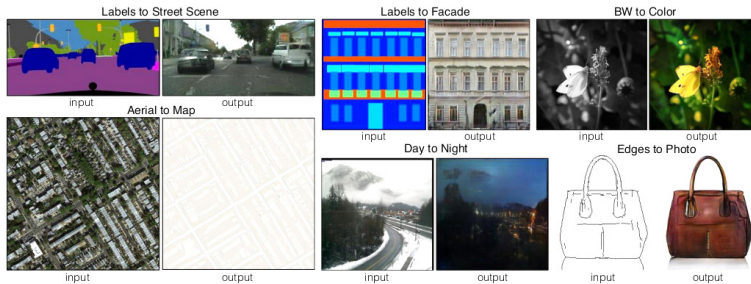
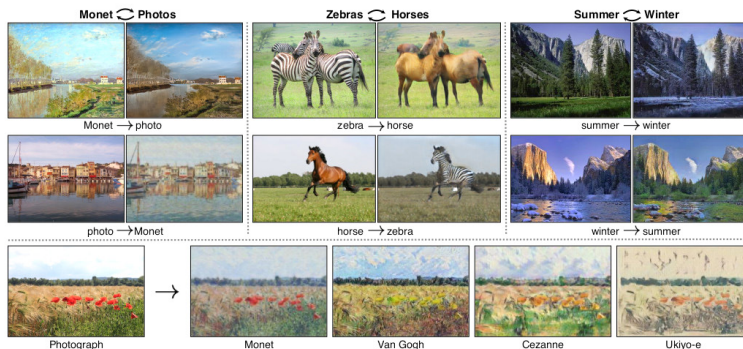


Image-to-image translation with cGANs (Isola et al. 2018)

CycleGAN

CycleGAN was developed by the same group as pix2pix, with a similar function. The improvement consisted of being able to train with unpaired images. If pix2pix required a great number of, say, outlines of bags with actual photos of bags, CycleGAN infers the characteristics of training sets such as the paintings of a given artist (Zhu et al. 2020). After the learning step, the system is able to apply specific styles to any image. The paper and code came out in 2017.



Unpaired image translation (Zhu et al. 2020)

⁷ <https://phillipi.github.io/pix2pix/>, accessed 28 August 2023

StyleGAN

The importance of this network rests on its capacity to produce high resolution photo-realistic images, on a quality that kept evolving across different versions. The first came out in 2019 (Karras, Laine and Aila 2019) as a development of 'Tero Karras' work on ProgressiveGANs (Karras et al. 2018). It became the tool of choice for websites that popularised photorealistic generative images such as This Person Does Not Exist⁸. It was also the system used for *Post-colonialism*, an artistic component of the thesis that will be seen a bit later in the text. A new version, StyleGAN3, was released in 2021 (Karras et al. 2021).

Super-resolution GAN

The relevance of super-resolution GANs derives from their ability to scale up low resolution images while keeping – or actually creating – detail. The technique is still used in the latest Stable Diffusion models, since their output always needs to be upscaled. ESRGAN came out in 2018 (Wang et al. 2018) while a more realistic model was released in 2021 (Wang et al. 2021). These networks are still used in current diffusion models for scaling.

VQGAN

The last GAN I will describe is VQGAN, which opened up the path to the text-to-image models being used now (Esser, Rombach and Ommer 2021). Its innovation consisted in applying Transformers to convolutional image generation. *Transformers* is an architecture developed in 2017 that is instrumental in the creation of large language models like GPT (Vaswani et al. 2017). The authors of VQGAN were able to generate images with much better quality by using this technique. Besides, they were able to guide the generation with poses and depth maps, which give much more control over the final result.

8 The original domain was taken over by the company Stability.ai. Other versions of the tool are available at <https://thisxdoesnotexist.com/>, accessed 20 June 2023

This was the model used by Katherine Crowson⁹ and Ryan Murdock¹⁰ to develop text-to-image models that would become very popular.

GAN art

Artists have been working on artificial intelligence since its onset. After the interest in cybernetic theory, others explored the field as an expressive subject in the 1970s, such as Naoko Tosa and Harold Cohen (Wilson 2002, chap. 7.6). But GANs provided a powerful, AI-based novel tool for image-making. It is impossible to determine a ‘first’ artistic experiment made with GANs, and is probably not so relevant – artistic innovations should not be treated as inventions, where the first proponent is the one who gets to patent the deed. But one of the most immediate reactions that can be mapped came from Robbie Barrat. His Art-DCGAN code was published in 2017. It downloaded hundreds of artworks from the internet in order to train a version of Soumith Chintala’s DCGAN.

By splitting the sets in categories like nudes, landscapes and portraits, he was able to generate 128×128 images that, while far from a perfect recreation, displayed a strong but hard to define resemblance to the styles they have been trained on. GAN-generated images of this period are easy to spot: they have a blending texture, where elements melt into each other, resulting from the convolutional techniques described in the last chapter. Depending on the artist, this was either incorporated into the work or was a problem to be fixed by working on better training and cherry-picking results.

The images made an impact on the artistic community. In the month after its publication, Hugo Caselles-Dupré, a member of the French artist collective Obvious, joined a thread¹¹ in Barrat’s repository asking for help to fix a bug. A few months later, in 2018, the collective created a collection of AI-generated portraits depicting the fictional *Belamy* family. One of these, named *Edouard de Belamy*, ended up being sold by the world-renowned auction house Christie’s for 432,000 US dollars in October of that year (Jones 2018).

9 <https://twitter.com/RiversHaveWings>, accessed 28 August 2023

10 <https://twitter.com/advadnoun>, accessed 28 August 2023

11 <https://github.com/robbiebarrat/art-DCGAN/issues/3>, accessed 28 August 2023

The story behind the sale is rich in controversies that bring to light the issues surrounding AI-generated art (Bailey 2018b, 2018a, 2018c). The three members of Obvious had no art training. The only member with computer training was Caselles-Dupré, who was the starting his PhD in machine learning. The trio tried to sell the artworks through Twitter and eBay without success, until Christie's approached them on Twitter offering to auction it. Even before the sale, the auction made ripples in the generative art community, as other artists had been experimenting with algorithmic art for a long time and felt despised. Moreover, it fuelled the argument that art-istry was not required to make computer-based artworks.

Questions of authorship also surfaced. At first sight, it seemed that Obvious had not only used Barrat's code, but also the models trained by him. This was denied by Caselles-Dupré, who posted a video proving they had done their own training. But the question of whether Barrat should have a share of the profits was raised: after all, Obvious' work was a direct implementation of his code. Caselles-Dupré dismissed the issue, stating that Barrat also fails to credit scientists like Goodfellow, who invented GANs. The issue is further complicated by the fact that Obvious chose to sign the painting with a mathematical formula from Goodfellow's paper, representing the rules of the game played by the generator and discriminator networks. In fact, the public relations material from Christie's and Obvious initially stated that the author was the AI system. This stance was later rectified in the interview used as reference here. Finally, *Belamy* can (and should, according to the group) be interpreted as a translation of Goodfellow's name into French (Bailey 2018a).

Barrat continued to develop GAN-based artwork. While Obvious was working on *Belamy*, he integrated AI into crypto-based art, giving away NFTs minted with 300 nudes from the nude GAN model. In 2019, he was contacted by a gallery in Paris to develop a project with painter Ronan Barrot. Catherin Vossen and Albertine Meunie, working for Vossen Gallery, realised that his obsessive paintings of skulls – hundreds of them – made them the perfect set for AI training. Their collaboration ended up in a show where Barrot's paintings appeared next to the skulls generated by Barrat. The results came out strikingly similar to the original artwork, and the painter said

he wished he had painted some of those himself. That led Barrat to create a new batch that would have some of his own input, by changing the training set enough so that he could make the machine ‘hallucinate a bit’ (Table ronde Infinite Skulls 2019). Vossen Gallery opened a new show¹² with his AI-generated work in 2023, with works from 2018 until 2020.

It would be no exaggeration to say there was a GAN-based art movement around the late 2010s that explored its peculiar visuals and generative features. François Chollet, an important artificial intelligence scientist, was probably the first to use the term GANism: ‘GANism (the specific look and feel of seemingly GAN-generated images) may yet become a significant modern art trend’ (Chollet 2017).

It bore the aesthetics of uniqueness in multiplicity, in ghostly representations (Waelder 2020). Several artists were lured by the technique and developed a practice around it. Helena Sarin experimented with the textures of GANs to develop a unique style. Anna Ridler generated a tableau of tulips¹³ as a reference to the economic flower bubble¹⁴ of the seventeenth century in the Netherlands. It also referred to one of the best-known collections in data science, the Iris dataset (originally published in the *Annals of Eugenics*, no less) (Fisher 1936). Sofia Crespo engraved the GAN look onto her biological-looking creatures¹⁵. Jukka Hautamäki worked with selfies¹⁶ to create monstrous versions of himself. Taller Estampa designed alien species¹⁷. Some works point in the direction of autonomous systems: Memo Akten proposed a network that would self-train¹⁸ itself on the images of a webcam in an installation. Mario Klingemann and Gene Kogan developed online, community-based art generators with strong claims to autonomy that will be reviewed later on.

12 <https://avant-galerie.com/robbie-barrat-2018-2022>, accessed 28 August 2023

13 <http://annaridler.com/myriad-tulips>, accessed 28 August 2023

14 https://en.wikipedia.org/wiki/Tulip_mania, accessed 28 August 2023

15 <https://neuralzoo.com/>, accessed 28 August 2023

16 <https://jukkahautamaki.com/2019/03/11/synthetic-selfies-2018/>, accessed 28 August 2023

17 <https://tallerestampa.com/estampa/especies-marcianes/>, accessed 28 August 2023

18 <https://www.memo.tv/works/learning-to-see/>, accessed 28 August 2023

Artbreeder¹⁹ is also an interesting project since it has its roots in a different branch of artificial intelligence research. Joel Lehman and Kenneth Stanley have been working on the concept of goalless computing – the idea that designing systems that do not have any specific objective might be a more efficient way to breed novelty (Lehman and Stanley 2011). To put their ideas into practice, they designed Picbreeder²⁰, a website that from 2007 until 2021 allowed users to aimlessly combine images into new hybrids according to their subjective evaluation of interestingness (Secretan et al. 2011). The inbreeding approach added evolutionary features into the generative processes, yielding interesting results, which quite often arrived at familiar shapes.

Inspired by this previous work, artist Joel Simon created a website in 2018 that integrated GANs into this concept, the GANBreeder²¹, using BigGANs. The project grew a strong community of users who shared their creations on the forum-like interface, and became the scenario of another intellectual property dispute around generative imaging (Bailey 2019; Zeilinger 2021a). Artist Alexander Reuben used images from GANBreeder to create his project amalGAN²². A complex creation process that included measuring the artist's brain waves used, in the last steps, Chinese for-hire artists to paint the images on canvas. But when the results were publicised, other GANBreeder users identified pictures of their own creation. Reuben claimed that he assumed the images were being created by an algorithm and were free to copy. However, even if they were 'bred' by human artists, the American copyright office says that, being machine generated, they might not be registrable (USCO 2021). Reuben admitted the misuse of images and addressed the issue by updating the system in a way other users' images would not be used (Bailey 2019).

19 <https://artbreeder.com>, accessed 28 August 2023

20 <https://web.archive.org/web/20071009011927/http://picbreeder.org/>, accessed 28 August 2023

21 <https://web.archive.org/web/20181119174641/https://ganbreeder.app/>, accessed 28 August 2023

22 <https://web.archive.org/web/20190215103728/https://areben.com/project/amalgan/>, accessed 28 August 2023

In 2019, Simon's project adopted the more generic name ArtBreeder²³. The technology it used kept up with developments in the field and, to this date, its community seems to be still going strong. But the list of GAN artists and creative projects is much larger than these samples and would justify a doctoral project of its own. In 2022, GANs would be abruptly dismissed by the generative community, as the more advanced *diffusion* methods dominated the field. But the few years when GANs was the method of choice were enough to leave an aesthetic imprint on generative arts. My personal artistic exploration of GANs was incorporated in the research and will be described soon.

Autopoietic aesthetic arrangements

The Barrat/Obvious episode around GANs is useful at this point to introduce some seminal concepts within this research. In a paper from 2021 I described how difficult the question of defining the art object for art and technology practitioners is, and how it gets entangled with the concept of inventions, especially in the face of societal paradigms of authorship such as patents and copyrights.

I will use these two abstractions to frame my argument, as they are the somewhat standardised ways society deals with different types of creativity. The patent office is available for the registration of new systems and ideas that produce an output, or an effect, and the copyright office is there to register the fixed form of creative outputs. To be eligible for a patent, an invention must have a utility. At the same time, copyright is only granted to the fixed output of an artistic process, and not to the idea behind it. Hence, there is a gap in dealing with the authorship of new creative arrangements — systems that are creative per se and do not necessarily fall within the utilitarian paradigm of a patent (Caldas Vianna 2022).

To help limit and define the concept I am trying to convey, I am borrowing the name of a practice from the music field, which has its own problems in solving authorship issues. An *arrangement* is the organisation of melodies and harmonies of different

23 <https://web.archive.org/web/20190825043841/http://www.artbreeder.com/>, accessed 28 August 2023

instruments for the performance of a song. This is an intensively creative process, which is not as recognised as song writing or performing. While it is easy to identify the composers and interpreters of a given piece, the creator of the arrangement is not so easy to locate. Some of the most well known melodies, such as bass lines, are the result of this process and might never be given appropriate recognition and will probably never appear on music scoresheets (Niles 2014).

The arrangement I propose here is a reframing of the creative process from a cybernetic perspective. It suggests a different point of view of dualities such as artwork/invention, artist/tinkerer and copyright/patent from the possibilities afforded by generative media.

Vilém Flusser describes the actions resulting from the machinic entanglement of humans and apparatuses as collaborations (Flusser 2011). However, he sees humans more as creative executors of the machine's programs, actively selecting from the algorithm output (Zylinska 2020, 53). This is why Flusser's philosophy stimulates the deconstruction of the black box, the closed systems running the algorithms. I see the current practitioners in art and technology going further and further beyond Flusser's exhortations and organising the black box itself. Most productions within art and technology require this organisation of software and hardware — used here in the broader sense of *instructions and tools* — to propose a new creative arrangement that might have outputs. To emphasise their role within arrangements, I propose to label these outputs as *by-products*. Very often the by-products of the arrangements are also artworks (images, sounds, music, text) that I consider secondary, for in this model I am suggesting, the main form of expression is the system itself. Depending on the weight of non-human agency within, by-products can be seen as the results of mere menial jobs in the arrangements.

Generative art is the essential *lieu* of the arrangement, as it can be seen as the art of the autonomous systems. One of the most quoted definitions of generative art says that it:

'[...] refers to any art practice in which the artist uses a system, such as a set of natural language rules, a computer program, a machine, or other procedural invention, that is set into motion

with some degree of autonomy, thereby contributing to or resulting in a completed work of art' (Galanter 2003).

When Michael Noll or Vera Molnar programmed systems that generated images in the 1970s, they were making arrangements that outputted by-products. When Casey Reas uses the *Processing* software to propose pictorial algorithms, he is carrying on this tradition.

One way to understand Walter Benjamin's concept of *aura* is to describe it as what gets lost in the technical reproduction process, that is, the lure of the original work, its *here and now*. But the problem now is no longer reproduction. Of course, one can make any number of copies from a given algorithmic print. The problem is that one painting can have endless variations that follow the same rules and thus proportionate a similar effect. Once the rules are in place, the number of possible instances of the same work is unlimited. Hence, within the framework of generative art, what withers is the output, the by-product, these possible generations created from the same arrangement. If we are to look for what Benjamin calls *authenticity* of the artwork, we must look into the ritual of crafting the system. 'The unique value of the "authentic" work of art has its basis in ritual, the location of its original use value. [...] The whole sphere of authenticity is outside technical—and, of course, not only technical—reproducibility' (Benjamin et al. 2008).

My work *Devorondina* (2010) is a mobile sensing unit that translates environmental conditions into image by-products. Jonah Brucker-Cohen's Camera Face²⁴ (2022) is an application that produces images of people who are not looking at a camera. Julius von Bismarck's *Image Fulgurator*²⁵, a photographic contraption of which by-products are photographs that reveal invisible projections, is one of the few examples where the artist actually holds a patent for the arrangement. Every generative art example seen so far in the research is an arrangement of software (or plain instructions, in the

24 <https://web.archive.org/web/20220525154729/http://www.coin-operated.com/2022/03/20/human-error-camera-face-2022/>, accessed 28 August 2023

25 https://web.archive.org/web/20200201000000*/http://juliusvonbismarck.com:80/bank/index.php/projects/image-fulgurator/2, accessed 28 August 2023

case of Sol Lewitt, Tristan Tzara and others), where the generated output is its by-product.

In some examples, the by-products disappear, merging into the arrangement itself. The fate of Ana Laura Cantera's bioconstruction autoenergetic bricks²⁶ is to become a wall. Gilberto Esparza's Nomadic Plants²⁷ arranges mechatronics and plants in a performative action. Tuula Närhinen's Storm in a Teacup²⁸, part of her Clapotis series, is a contraption consisting of a hose submerged in the ocean, and a tube that reveals the tidal movements by the level of seawater against the horizon (Närhinen 2022).

The concept is not exclusive to contemporary art and technology works, as ever since the appearance of art it has somehow implicated the use of tools. Therefore, oil paintings are by-products of the arrangement of brushes, oil, canvas and frames; books result from arrangements of language coded in textual signs that in some instances require sheets of paper bound in volumes. It may sound even offensive to consider, say, Bach's *Matthaus Passion* as a by-product of music arrangements of the Baroque. But this displacement is necessary to understand the shift recently afforded by generative creation.

Cinema takes a complex and well-established arrangement of equipment, writers, crew and sensitive material into a by-product to be enjoyed in dark cinemas, on TV sets and, recently, mobile screens. But when a filmmaker like Marcelo Pedroso asks for the tourists on a sea cruise to give him the images produced during the trip to make them into a documentary feature (Pacific, 2009), he is proposing a new cinematic arrangement. In my work *Ressaca* (2008), each cinema session was a by-product of the careful arrangement of sequences created by means of a specific software in a performative environment.

26 https://web.archive.org/web/20230420105129/http://nuvem.tk/wiki/index.php/Ana_Laura_Cantera, accessed 28 August 2023

27 <https://web.archive.org/web/20230420104811/http://gilbertoesparza.net/portfolio/plantas-nomadas/>, accessed 28 August 2023

28 <https://www.tuulanarhinen.net/artworks/clapotis/storm.html>, accessed 28 August 2023

The model is made clearer when the arrangement takes importance away from its by-products, as their post-humanist appeal questions human authorship and takes over their agency into an agency of the artificial. That is what leads us back to the GAN art examples. GANs themselves entail an arrangement in their combination of generator and discriminator networks to produce outputs. From this starting point, Barrat proposed a new configuration that consisted of feeding downloaded artworks, organised in categories, to create new artworks. *Obvious*, in turn, used Barrat's arrangement with minimal modification to create the portraits of the fictional Belamy family.

Arrangements are prone to simultaneous discovery, the phenomenon in science and engineering where something is invented or found by different researchers without knowledge of each other's work. It happened to cinema technologies, aeroplanes and others contraptions. As described in my aforementioned paper, Brazilian artist Pedro Veneroso and British duo Random International developed the same installation unbeknown to each other, consisting of a controlled rain room in which the dripping was interrupted where visitors passed. This happens because, in my scheme, they belong to the domains of concepts, and are the effect of their worldly zeitgeist, or their common métier, or unconsciously shared questionings. That does not mean they are the same artworks: the arrangement allows for a variety of instantiations of the concepts, which with a unique materialisation, circulation and reception.

This type of evolutionary path is a particularity of arrangements. In open collaborative software, the act of duplicating a set of programming code for improvement is called a fork. During a fork, a duplicate of the original repository is made, and after that the forker is free to modify the copy at their will. Arrangements are prone to forks, since they are conceptual constructions. Had they used the complete software and training data from Barrat, *Obvious* would be creating just a copy of his arrangement. But adapting the training data to their own curation after the duplication is a change that, while minimal, could probably justify its presentation as a new work — a fork.

To define the limits of the arrangement, I will return to the term *autopoiesis* from biological cybernetics. 'What defines (living systems) is their autopoietic organization, and it is in this autopoietic

organization that they become real and specify themselves at the same time' (Hayles 1999, 151; Maturana and Varela 1980, 48).²⁹ Arrangements are circular in that their components produce effects that define the arrangement in itself, and allow it to maintain its productive condition. 'The living organization is a circular organization which secures the production or maintenance of the components that specify it in such a manner that the product of their functioning is the very same organization that produces them' (Maturana and Varela 1980, 9).

'An autopoietic system produces itself while simultaneously producing its own conditions, both internal and external' (Keenan 2022). The autopoietic condition of arrangements presupposes that their organisation, which is built around self-maintenance, defines their limits and the category they belong to.

In this framework, a system that entails an output different from itself and is contingent on an external observer is not autopoietic, but *allopoietic*. An example given by them is the human organisation of car factories, which produce cars and not other factories. But an *allopoietic* system also bears autopoietic qualities; the factory must be able to fix itself in order to preserve its productive arrangement. In my own fork of their framework — the duplication of its building concepts for further modification, described above — I will not create a binary distinction between these categories, but a gradual transition. Human-made completely autopoietic systems exist only in the imagination. Arrangements are never completely autopoietic, and will always have allopoietic features. As the Chilean biologists' theory states, allopoietic machines are not autonomous, but contingent on their observers and their goal, which lie beyond their closure. That is the case of even arrangements that display a high amount of agency (Maturana and Varela 1980, 80). Generative art is the quintessential *autopoietic* art, as it relies on systems whose arrangements define the systems themselves. It thrives precisely when located at the boundary between its output and its autonomous features.

29 I could not find Hayles' quotes of *Autopoiesis and Cognition* either in the original edition or the edition of the Maturana and Varela book she refers to. I leave here both references for clarification.

And here appears one last element of the concept. It is hard to define agency. It has been appropriated by fields as varied as psychology, neuroscience, and even narrative studies. We will revisit the concept later, when venturing around issues of consciousness. Particularly for this concept, let us see it as an original impulse that appears in the act of inventing the arrangement, in the production of their by-products, and is distributed in the processes of self-maintenance.

This definition is limiting and does not consider animals, for instance, as having creative agency. The originating impulse is human agency, and arrangements cannot escape this condition yet — at least while there are no examples of non-human origination in Lovelace's sense. But the human arranger also defines, within the system, how agency should be distributed, and how much agency they are willing to give up. When they forsake their own agency, the arranger opens up room for environmental agency (through sensors), participatory agency (interactive pieces), mechanical or random agencies and other types.

Arrangements where the author relinquishes a considerable amount of agency to the system lean strongly into a post-human condition. In every artwork that is exhibited, the artists necessarily share some of the agency with their audience, since they have no control over how the piece will be received, interpreted or felt (Lebel et al. 2021, 77–78). In the case of interactive works, this surrender is even more explicit. But arrangements convey a configuration where human protagonism is displaced and intelligence is co-produced with machines, tools and computers, echoing the framework proposed by N. Katherine Hayles. In her account, 'distributed cognition replaces autonomous will; [...]; and a dynamic partnership between humans and intelligent machines replaces the liberal humanist subject's manifest destiny to dominate and control nature' (Hayles 1999, 288). Expanding on Edward Hutchins' account of cognition (Hutchins 2006), Hayles also refers to Searle's Chinese Room experiment to state that the cognition doesn't dwell exclusively in the human inside it, but in the box as a whole. 'The situation of modern humans is akin to that of (the human) in the Chinese room, for every day we participate in systems whose total cognitive capacity exceeds our individual knowledge' (Hayles 1999, 289). When

we engage with agential autopoietic arrangements as authors, viewers and programmers, we grant some cognition and agency to the arrangement as a system, as well as the outer system of relations where it is also embedded.

In this view, post-humanism is a vector that points away from human individualism and the romantic idea of the author. But outside non-human domains — mechanic or animalist — the author subsists, as there is no creative arrangement organised exclusively by another species or a machine. While it is true that the individual's role is less important every time, the point where it disappears has not been reached.

A hermeneutic process crosses through the arrangement, from human to extra human, to the outside, to the thing, to the creation. Autopoietic arrangements imply authorship. Hermeneutics reifies itself as the allopoietic component of the arrangements. They are forms of creative expression, which quite often are not visible or not recognised as such. It is also a concept that opens a vector towards a broadening of expressive authorship to inventors or tinkerers who do not see themselves as artists.

In the rest of this text, every time I use the term *arrangement*, I am referring to this agential aesthetic autopoietic arrangement concept, which will become clear as we refine its meaning and see more examples.

Artistic component: Post-colonialism

Post-colonialism is one of the pieces displayed at the *Singular* exhibition, a collective show which was part of the Research Pavilion 2021 edition at the Hietsu Pavilion in Helsinki, and included other research works of mine described here. The experiment was not focused on GAN's rummy graphics, but on issues of bias, uniqueness and multiplicity. In fact, the algorithm of choice was StyleGAN2, which at the time was known to produce the most photorealistic images.

The work took the shape of a triptych, with three digital prints measuring 100x100 cm each. The central piece has 10,000 human faces generated using StyleGAN2. The model of choice was the

Flickr-faces³⁰ dataset, which was in fact created for the first version of StyleGAN. After the issues raised with ImageNet, it is no wonder this collection was organised with plenty of safeguards and a fine attention towards privacy and content. According to the documentation, ‘Only images under permissive licenses were collected. Various automatic filters were used to prune the set, and finally Amazon Mechanical Turk was used to remove the occasional statues, paintings, or photos of photos’ (NVlabs/ffhq-dataset 2023). A tool is included to check if a given Flickr user account has photos in the set, and the institution provides a method in case people want their pictures removed. Lastly, a warning states that the set should not be used for ‘development or improvement of facial recognition technologies’.

But no matter how careful the collecting is, the set cannot be better than the data being gathered. The distribution of Flickr users – and, in consequence, of the photos they upload – is heavily biased towards Western, computer-educated people who have the know-how to store their photographs on the platform, and to select a permissive licence. The result is that it is in all likelihood composed for the most part of photos of white people.

To explore this hypothesis, two more panels were made starting from the same 10,000 pictures. For the first, I wanted to use an automated service to identify the ethnicity of the faces. Unfortunately or not, by that time commercial providers of AI visual analysis had made such services unavailable. As such, I looked for a repository of open software that could do the same task, and found one written by a machine learning engineer, Donghyeon Won (Won 2023). To train his model, he used a dataset from the Chinese University of Hong Kong, created for the same purpose of learning face attributes (Liu et al. 2015). The paper claims that the annotation of the faces, downloaded from the web, was made by a professional labelling company.

I adapted this software to process my own StyleGAN tableaux (Caldas Vienna 2023). The only difference from the original repository is a script to process the 10,000 images. I used the system to classify the faces into different ethnicities and removed the ones

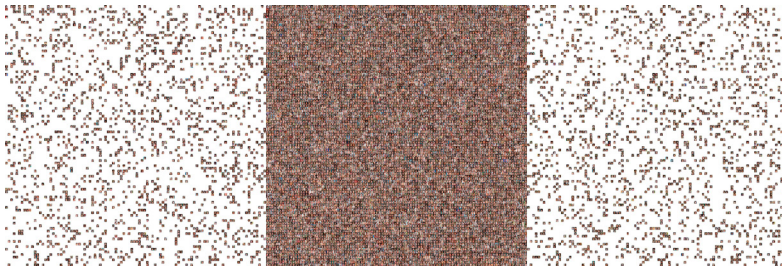
30 <https://github.com/NVlabs/ffhq-dataset>, accessed 28 August 2023

automatically tagged as white. Since they made up most of the images, I ended up with a tableau with large areas of white space. Then I proceeded to do the same operation, but manually – by looking at each face and deciding if they were white ‘enough’. By doing this, I immersed myself in my own judgments, realising how much whiteness can also be defined by factors other than skin colour. For instance, a man of mixed race could have the scales tipped towards white if he wore well-dressed suits.

To frame it within my own concept, the arrangement of *Post-colonialism* is made of the forked repositories, my own adaptation which hooks up the ethnic classification to StyleGAN-generated faces, the triptych as a disposition option, and my act of classification. The tableaux themselves are by-products, in the sense that any combination of faces and white space would serve the discursive purpose of the piece.

Post-colonialism is a comment on the cultural space shaped by hegemonies. The work takes a stance by appropriating the problematic uses of AI to turn the tables against a colonial representation. The white space is a testimony to epistemic violence, the bleaching of diversity in artificial intelligence systems that refer to the underlying world that they intend to model.

No statistical account of the results is presented, as the panels themselves should be a visual representation of the biases of the datasets. More images can be found at the Research Catalogue exposition.³¹



Post-colonialism, digital print, 300 x 100 cm, 2021

31 <https://www.researchcatalogue.net/view/1408828/1408855>, accessed 28 August 2023

Visual AI's own linguistic turn

Three developments in AI served to push generative visual techniques away from the training paradigm, which required hundreds of similar images, and into the text-based, descriptive prompt. With this claim, I do not refer to the generic model creation techniques, but to the practice of artists working with visual AI. For them, the days of GAN implied curating a particular collection of images that would be operationalised into the artwork/model. Broadly speaking, creators using today's text-to-images focus on *prompting* techniques using public models, rather than the in-house training of neural networks.

In any case, I identify these developments as VQGAN, as already described, OpenCLIP and finally Stable Diffusion. However, the effort to create images from text descriptions started earlier than this.

In 2015, a new technique allowed the generation of images that depicted sequences of digits. But it was later in the same year that a paper demonstrated what were probably the first recognisable images built from text captions (Mansimov et al. 2016). It used a Variational Autoencoder (VAE) to produce blurry images that were then sharpened by a GAN, a solution similar to Stable Diffusion's final upscaling. Another approach published in the next year yielded results with much more definition (Reed et al. 2016). It used GANs as generators, and although they tested the generic image set MS-COCO, which contains 91 categories, the results seemed much better from specialised sets, like the Caltech database of birds or the Oxford-102 flower set (Lin et al. 2015; Nilsback and Zisserman 2008; Welinder et al. 2011).

Other text-to-image systems followed, like *StackGAN*, *StackGAN++*, *AttnGAN*, *Obj-GAN* and *TReCS* (Koh et al. 2021; Wenbo Li et al. 2019; Xu et al. 2017; H. Zhang et al. 2017, 2018) But the one that attracted most interest for its quality and availability was OpenAI's Dall·e³². It was announced in January 2021, with the code and paper being published in February (Ramesh 2023; Ramesh et al. 2021). A dataset made of 250 million image pairs was scraped — the term used to describe massive hoarding of public internet material.

32 <https://web.archive.org/web/20210105192048/https://openai.com/blog/dall-e/>, accessed 28 August 2023

The training was done through a discrete variational autoencoder (dVAE) to compress images into a 32×32 grid of image tokens, resulting in a vector with a length of 1024 numbers. Each image token can have 8192 values. This vector is concatenated with 256 tokens of text encoding, which is then used to train a Transformer.

This first version of Dall-e could already spawn convincing images from a broad domain of text prompts. It attracted great interest from the generative AI crowd and the media in general. But the real push towards text-to-image would come from another OpenAI release. OpenCLIP³³ was announced the same day as Dall-e, and it is curious that the company did not use it to power the text-to-image platform, since it became the *de facto* standard for such systems in the following years (Radford et al. 2021).

It uses the same idea described above: image and text embeddings are trained together as one long vector. Given any text and image, the model can output the probability that the text describes the image correctly and vice versa. The fact that CLIP was released as open source was instrumental in its immense popularity. Two weeks after it was published, there was already code published³⁴ by Ryan Murdock, who hooked it up to a BigGAN generator (Brock, Donahue and Simonyan 2019). On the same day, Eyal Gruss³⁵ published his adaptation³⁶, and many others followed. Katherine Crowson was probably the first to connect CLIP to a VQGAN, in April: ‘encode style with CLIP, encode content with VQGAN. in a loop, decode current VQGAN latent, encode it with CLIP, compare result with style, obtain gradient for VQGAN latent, update VQGAN latent’ (Rivers Have Wings [@RiversHaveWings] 2021).

33 <https://web.archive.org/web/20210105192307/https://openai.com/blog/clip/>, accessed 28 August 2023

34 <https://twitter.com/advadnoun/status/1351038053033406468>, accessed 28 August 2023

35 <https://twitter.com/eyaler>, accessed 28 August 2023

36 https://github.com/eyaler/clip_biggan/blob/3e131c799bb055798ca3f861344a52e587432921/ClipBigGAN.ipynb, accessed 28 August 2023

The code is probably the one published in this interactive programming notebook³⁷, although it cannot be confirmed since such notebooks are not timestamped. The corresponding paper was only published in 2022 (Crowson et al. 2022). Keeping track of all these innovations is not easy, so let us just bear in mind that VQGAN had been released only a few months earlier (Esser, Rombach and Ommer 2021). Soon after that, commercial sites offering text-to-image renders started to appear. NightCafe³⁸ would be launched sometime in May 2022, selling AI-generated prints and a tool to create them. StarryAI³⁹ launched in August 2022.

But VQGANs have some aesthetic limitations. Some cherry-picked results may look fantastic, but most of the time the images show the uncanny blending characteristic of GANs. By the end of 2022, the community started to look into a solution that had first come out in a paper from 2015. Jascha Sohl-Dickstein, a researcher at Stanford University, had proposed to look back into non-equilibrium thermodynamics, one of the most ancient concerns within systems theory (Sohl-Dickstein et al. 2015). The concept of entropy can be explained as a measurement of the degree of disorder in a system. If this system is an image, we can think of the disorder as visual noise, or in generic data, the diffusion. Entropy can also be seen as an arrow of time: events that increase entropy are only reversible if energy is put into reversing the disorder.

One of the concerns of systems theory is to explain how spontaneous order appears in natural processes. Life is an example of the self-organisation of chemical elements to maintain and reproduce the living system.

The innovation in Sohl-Dickstein's paper is to use a formula that removes noise from data in order to reverse the diffusion process. Besides whole images, it showed how this could be used for *inpainting*, that is, redrawing parts of the frame that have been lost or

37 <https://colab.research.google.com/drive/15UwYDsnNeldJFHJ9NdgYBYeo6xPmSelP>, accessed 28 August 2023

38 <https://web.archive.org/web/20200519133524/https://nightcafe.studio/>, accessed 28 August 2023

39 <https://web.archive.org/web/20210826024415/https://www.starryai.com/>, accessed 28 August 2023

masked. Text-to-image diffusion appropriates this idea in a process guided by the embeddings obtained by the text prompt to create a sharp picture.

By the end of the year, several independent programmers were tinkering with Diffusion and OpenCLIP. It is difficult to make an archaeology of the process, since most of the development seems to have been done on notebooks in Google COLAB virtual machines, which save neither timestamps nor a history of changes. In any case, all the notebooks I could find state that the software named Disco Diffusion originated, again, in a notebook by Katherine Crowson. Versions three⁴⁰ and four⁴¹ can still be found, with credit to programmers who contributed with different features, like Daniel Russell⁴² and Chigozie Nri⁴³. In January 2022, a repository⁴⁴ was created, with the latest version by Max Ingham⁴⁵.

Experiments with diffusion at OpenAI were published in February (Nichol 2023; Nichol and Dhariwal 2021). In June, it was clear that this process would produce better results than VQGANs: ‘Diffusion Models beat GANs on Image Synthesis’ (Dhariwal and Nichol 2021). In December, the company published GLIDE, which combined OpenCLIP with image diffusion, with the corresponding paper coming out in March 2022 (Nichol 2021; Nichol et al. 2022).

In April, the second version of Dall-e was released⁴⁶. It incorporated Diffusion technology (Ramesh et al. 2022). In May it was Google’s turn to publish their diffusion-based text-to-image platform, Imagen (Saharia et al. 2022).

CompVis, the same research group from the Ludwig Maximilian University of Munich that released VQGAN, developed a diffusion method that trained the models in latent space, which is much more

40 <https://colab.research.google.com/drive/1bItz4NdhAPHg5-u87KcH-MmJZjK-XqHN#scrollTo=BGBzhk3dpcGO>, accessed 28 August 2023

41 https://colab.research.google.com/drive/1DHvr-DSNtlpvf0HgI0QmCLSe_PtShzPH?usp=sharing#scrollTo=5ui9Pco33nCN, accessed 28 August 2023

42 <https://github.com/russelldc>, accessed 28 August 2023

43 <https://github.com/ChigozieNri>, accessed 28 August 2023

44 <https://github.com/alembics/disco-diffusion>, accessed 28 August 2023

45 https://twitter.com/Somnai_dreams, accessed 28 August 2023

46 <https://web.archive.org/web/20220406141041/https://openai.com/dall-e-2/>, accessed 28 August 2023

compact than the image space, allowing faster processing time (Rombach et al. 2022). This paper used the Laion 1.45B dataset, a collection of web-downloaded images 100 times larger than ImageNet. Laion is a non-profit initiative run by AI enthusiasts, with an interesting role in the aesthetics of AI that will be discussed in the next section. For now, it is important to know that they also created the Laion-Aesthetics dataset, a 400 million image subset of Laion-5B filtered by an automated predictor of visual quality.

The latent model trained with this dataset was the basis for Stable Diffusion, the platform that became a standard for text-to-image synthesis. Laion received support for training the model from the company Stability AI ⁴⁷, who then adopted it for its commercial synthesis product, DreamStudio. It was also adopted by MidJourney ⁴⁸ in 2022, although they later started to train their own models.

Stable Diffusion impresses for the realism of the images it generates. Several control methods are being developed by independent programmers and added to the open pool of resources around the model. This is giving creators more and more control of the results, like defining poses and replacing undesired artefacts. From being bounded to a specific domain – faces, cats, churches – generative imaging became limited only by what can be expressed linguistically (at least in principle, since some concepts are still hard to be translated to visual representations). Most impressively, the technology is not the result of a secretive closed corporation, but of a highly entangled ecosystem spanning from non-profit institutions, university researchers, artists, independent programmers, and tweekers of machine learning models, and even companies committed to open sourcing its developments. Without this open collaborative environment, the development of text-to-image systems would not have been so vertiginous.

The generative algorithmic image not based on machine learning already bears a complexity from its underlying text: code, data and knowledge, described by Nake as the *subface* of the image (Lee 2018). With text-to-image models relying on vast collections of unaffiliated pictures, this complexity increases exponentially. To

47 <https://stability.ai/>, accessed 28 August 2023

48 <https://www.midjourney.com/>, accessed 28 August 2023

describe these generative systems, I prefer the term *semantic imagining*, mainly to stress their ground-breaking ability to translate meanings. Independently of how it might be named, the process is the manifestation of an extreme case of distributed cognition, in the sense that meaning and agency are spread over a variety of actors like human-made concepts (like code or prompts), data sets, trained neural models and society.

ARTIFICIAL GENERAL INTELLIGENCE

By now we have seen the evolution of AI-generated imagery from the early strange-looking results of training GANs, to beautiful translations of prompts into images. We have looked into the issues raised by the methods they use, namely the fact that in order to output new content, they need to learn from millions of existing images that were created by humans using manual or automated tools.

We have also seen how the advancements in computer-generated content brought us closer to the ideal, unreachable autonomy of the art machine. In this section of the project, I'd like to focus on limitations and potentials of artificial intelligence with regard to this feature. How many of the results obtained currently by image synthesis models can surprise users by providing meanings and sensations that go beyond what they are trying to convey? How does AI deal with ambiguity, metaphors, analogies? Disobedience is a human behaviour that lies outside machine's cognition — that I propose to be seen as a valuable innate skill. It was indispensable for the evolution of Western art. By delving into these questions, new creative arrangements between human and machine can arise, possibly leaning more and more towards a subjective agency from the side of the algorithm.

Referring to the previous chapters, many examples of hybrid or non-human art systems can be found. From the mechanisms of Hero of Alexandria to cybernetic art, there was a constant quest for creating ways to replicate human labour in creativity and entertainment. Mechanisms for mechanically reproducing artworks abound since the industrial revolution — lithography, cinema, photography, turntables. Most recently, digital synthesis and manipulation have also become powerful tools, allowing the simulation of musical instruments, movie sets and even actors.

The ultimate art synthesis platforms are today's machine learning systems. My project aims to demonstrate how and why they afforded a leap towards autonomous automated creativity for some

reason which I will explore, related to unseen uncontrolled uncertainty added to the generative algorithms, and to the way they fit within societal relations. But before that, I need to clarify what I mean by autonomous creativity and its limitations.

The concept of autonomy in organisms and other entities has been incredibly hard to grasp, even after decades of efforts from the field of systems theory. Recent insights from neocybernetics point to the difficulties in delimiting any sort of boundary between a system and its environment, risking the very possibility of even defining an organism as an element separate from its surroundings (Clarke and Hansen 2009, 116). Bateson already noted this hurdle when he stated that the unit of survival is a combination of the organism and the environment (Bateson 2000). To propose an autonomous entity that spawns artworks is even more questionable, as art is mostly defined by its relations within society (Zylinska 2020). These relations are not only an effect of exhibiting the artwork, but also form a feedback loop with the result of the influences from society and previous pieces on the creator.

It is no surprise that authors have applied cybernetical tools in politics in quests for autonomy. Sciences of self-organisation should suggest clear ways to foster autonomous practices at the level of the individual and the group. 'Cybernetics, the science of control and communication systems throws valuable light on the anarchist conception of complex self-organising systems' (Ward 1988, 50).

The idea of applying cybernetics to political sciences surfaced in the early days of cybernetics and continues to be explored to this day (McEwan 1987; Sam 1979). As recently as 2021, Swann proposes to understand autonomy under clear categories. Political autonomy (as the autonomy *from*, or independence), is split into individual and collective, and we can file, for instance, autonomous regions into the latter one. Functional autonomy (as the autonomy *to*) is divided into tactical (how to get the job done) and strategic (what needs to get done) (Swann 2021). Nunes noted the organisation of social movements relates to the nodes of neural networks (Nunes 2017). The appearance of the concept of the DAO (decentralised autonomous organisation), connected to blockchain technology and cryptocurrency cultures, also inspired researchers to look into such references (Zargham et al. 2023).

Autonomy appears within the arts in the literature of Theodor Adorno, as a resistant force against culture industry (Adorno, Adorno and Tiedeman 1997, chap. 2). He builds on Immanuel Kant's notion of autonomy with a view of the art object and its capacity to stand on its own, not relying on context or even the artist. For Kant, fine art is 'a kind of representation that is purposive in itself and, though without an end, nevertheless promotes the cultivation of the mental powers for sociable communication' (Kant 2000, sec. 44,46, p. 185). This defining passage has nonetheless been the subject of much debate on whether it should be construed as a defence that artworks should be isolated entities with no further purposes than art itself (Halsall 2008, 139; Haskins 1990).

In a public discussion regarding creativity in AI, the philosopher Sean Kelly argued that art itself is a social construct (Kelly 2019). As such, the intrinsic quality of drawings made by an imaginary painter robot is not so important. It is society — or the market, or the art world, or a social network, or all of them — that determines what is amazing and what is not. Since I don't expect such an autonomous creative robot to be built, all I can do is evaluate how autonomous the projects that attempt to build it really are. My project dwells on this fragile scale and on the small steps some artificial systems have achieved in producing art.

The piece named *Photosynthesis*, developed in the context of this research and described in a previous section, displayed some of these steps. It is able to draw original and unique trees — the *by-product* of this particular *arrangement* — continuously, *ad infinitum*, as long as it is exposed directly to sunlight. Eventually, though, it will malfunction, and it will be unable to fix itself. As technology evolves and projects get more sophisticated, it is not unlikely that such a device will eventually appear. But for the moment, this only exists in my own speculation.

To support my claims and substantiate the argument, I propose to imagine an autonomous automated artist, which I will call Darci. Darci might or not appear in the near future. It is a thought experiment, one standard against which I will compare current developments. Its features will be outlined from now on, as some key concepts are elucidated.

Darci is not autonomous in the sense that biological animals are. It cannot fend for itself, like produce its own electricity or repair its parts. In this sense, it is far from being autopoietic. It is also not the spawn of antecedent Darcis, but it is created by humans. It does not long for the political autonomy of cybernetic anarchism, although it is likely to emerge as a consequence of its creative autonomy. But what I propose for Darci is much simpler on one level, and almost impossible on others: it should be able to create art that provokes a similar impact on society that an artist would — a noticeable effect.

Yuk Hui describes how cognitive scientist Brian Cantwell Smith assessed the development of AI in a way that resonates with this concept: ‘The intelligent agent must be able to situate itself within the world, and in doing this, must also recursively engage and modify the world. The agent and the world must constitute a structural coupling that is not only biological but also semantic’ (Hui 2021, 241; B. C. Smith 2019). This coupling is borrowed — against, outside the context of cognitive biology and therefore perhaps inappropriately — from the concept of enactment. ‘[...] cognition is not the representation of a pre-given world by a pre-given mind but is rather the *enactment* of a world and a mind on the basis of a history of the variety of actions that a being in the world performs’ (Varela, Thompson and Rosch 2016, 9). This coupling produces meaning as it occurs inwards the system, as the AI models are trained on millions of cultural references, as well as outwards, as their production resonates within the world.

Darci’s production must not be a rearrangement of existing ideas, but something new. All these concepts are really hard to limit: artists are not raised in caves, beyond contact with previous culture. Therefore, there is always some kind of re-enactment, even by the most original geniuses.

Scientist Margaret Boden, who has written extensively on AI, suggests the following set of criteria for creativity: novelty, originality and unexpectedness. I will deal with the last requirement later on in the text, when discussing the concept of surprise. In Boden’s concept, unexpectedness presupposes new arrangements of ideas across different fields. It is one of the expected capabilities of Darci. Novelty is a given in the case of computers, which can be designed to output unseen combinations starting from a programmed set.

Originality, in her terms, is the potential to explore a known field and find novel outcomes within it, which is something that machine learning models are particularly useful for. They posit their aesthetic latent space at the fingertips of creators and the machines, laying it open for explorations and reconfigurations.

Another model of creativity, now from the literature of psychology, proposed four different types of the concept, namely the four Cs: (Kaufman and Beghetto 2009)

- Mini-C would be the personal type of creativity that is related to learning processes - ‘meaningful interpretations of experiences, actions and insights’
- Little-C would be the universal human gift of dealing with everyday problems and creative self-expression
- Pro-C is the creativity used professionally, whether required by trade or not, that does not bring special eminence outside their field.
- Big-C is the ground-breaking type, capable of making original contributions with the potential to change the field or society.

Just like the different types of creativity, the imaginary autonomous artist would only have an impact on society if it fits in the Big-C category. And as we shall see, it is difficult to make the argument that even the best generative tools do not reach above the Mini-C level. In other words, they might be good at learning but not at breaking from the learnt domains.

The goal of this doctoral thesis is not to build Darci or initiatives with the potential to perform as fully creative and autonomous artists. This would require expanding the scope of the research to include concepts of consciousness and artificial general intelligence, as we will see. Nonetheless, it is important to visit some important discussions on these issues, as a way of grasping the limitations of the systems being built today.

As seen in the introduction, the concern regarding the ability of machines to generate ‘original’ things was born together with the first computer programmer, Ada Lovelace. Her comment on the impossibility of the analytical engine to ‘originate anything’ is in line with my own conjectures on the impossibility of a machine that can disobey. Computers are tools, much like scythes or cars, which

are designed to perform certain tasks. When a software generates a compelling image, it does so because some humans engineered it and instructed it to do so.

Human artists are different. Among the most admired traits in art is the capacity to turn against the expected, the settled, the unwritten code. Arguably, Western art evolved in dialectical opposition to its own previous standards (Hui 2021, 31). The Renaissance perspective was a denial of medieval flatness. Impressionism denied the literal representations of new-born photography (and took the studio to the streets, much like photographers did). Then it was the Cubists' turn to reject natural perspective, and so on.

Examples of transgressive art abound. Artists often intentionally break the law, sometimes as a means of expression, sometimes as a consequence of what they want to achieve. Activist art frequently relies on trespassing or subjecting itself to censorship in order to draw attention to issues. Graffiti artworks are for the most part crafted on private or publicly owned property, forcing graffiti artists to be in permanent conflict with security enforcement. The point being made is not that disruption is indispensable to the evolution of art: other cultures might not rely so much on creative ruptures. Chinese art, for example, places emphasis on inheritance and preservation (Hui 2021, 31). But the effects of disobedience are visible, and without it the world would look very different today. Engineering a well-behaved autonomous system to make art would be meaningless when faced with more subversive talent.

The definition of a computer — or a Turing-complete machine — is a device that performs behavioural patterns according to the symbols laid down on a tape (Turing 1937). Not following these rules would make it stop working. It is against its nature and purpose. We tend to wonder about and fear the idea of misbehaving machines because of the human tendency to anthropomorphise them, by projecting our own disposition to defy written rules. To give this feature to a program would be to give it agency, which is a human, if not also biological, trait (Brembs 2010).

I will deviate momentarily from the argument to explore the idea of adding rule-changing abilities to the program itself. An acclaimed video game named *Baba is You* allows players to do exactly that. The rules of this game, in plain English, can be seen as the user is playing.

‘Wall is stop’ would mean that the characters cannot traverse walls. But the characters are able to move the words around and when that sentence is modified by removing the word ‘stop’, the character is free to go through walls. ‘Key is you’ would make the controlled character be a key instead of the standard sheep/dog protagonist Baba. Of course, this implies just that another layer of laws was created by the programmer — a code of codes.

This is similar to Russell’s paradox, which appears in his and Whitehead’s *Principia Mathematica*, and was later disproved by Gödel, as discussed before. Basically, a formal system cannot be defined by its own rules, or it becomes paradoxical. An external, independently defined set of rules, representing a higher hierarchy, is necessary for its consistency. Another way of visualising this paradox was proposed by Cretan philosopher Epimenides: ‘All Cretans are liars’, that is: if all Cretans are liars, this statement by a Cretan is also a lie. On the other hand, if the statement is true, then Epimenides is not a liar and therefore not all Cretans are liars. One way to solve this contradiction would be to have a non-Cretan to declare the same thing, making the paradox disappear. In other words, an externally defined rule would be necessary.

One anecdote from the domain of science fiction might help illustrate that artificial constructs, when based on self-contained rules, will eventually become paradoxical. Isaac Asimov laid down three laws of robotics in the short story *Runaround* (Asimov 1950).

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

The goal of the laws is to promote a peaceful coexistence between humans and robots, while letting humankind reap the benefits of automation. This story, however, is written to demonstrate the limits — or, to some degree, the uselessness — of systems abiding strictly to internal logical rules. It also serves to show how the fact of bestowing rules to an artificial being makes their autonomy

at times impractical, at times impossible. *Speedy*, the robot in this tale, is sent on an extractive mission to Mercury: it must obtain selenium from a region that is too dangerous for humans. When the engineers notice it hasn't returned from the task, they investigate what happened, and find Speedy stuck in a loop. As it tries to obtain the element (following the second law), it approaches a dangerous zone. Realising the dangers, it then turns away, abiding the third law. As it reaches a safer place, the second law takes precedence, and the cycle starts again. To break it from its spell, the mission leader puts himself in danger. Since the first law precedes over the other two, Speedy rescues him, exits the loop, and is able to return to the base. In this metaphor, the external intervention of the astronaut would represent the necessary *ex-machina* axioms of mathematics.

Therefore, one possible reason why humans and other life forms thrive is that they are not limited to following rules. For instance, the rules of survival in nature include procreation and self-preservation; however, humans often opt not to reproduce and may even decide to purposely end their own life. Of course, I am falling again into the trap of projection, this time supposing that a human brain has hard-wired instructions, like a code-abiding computer, when they belong to completely opposite domains. Besides the obvious material distinctions — carbon-based entities against silicon and metal contraptions — there is a major difference in the way information is encoded. Biological neurons are analogic in the sense that the signals received and emitted from them have a wide, continuous range of values. Computers, in opposition, use the binary system, where all information is stored as atomic units of zeroes and ones, true or false. With only these two possible states, computers have the capacity to preserve enormous amounts of data and perform any calculation that can be transcribed into symbols. They can even organise the analogue information that constitutes the simulation of neural networks powering the recent developments in AI.

I will veer off from the argument again to note that the success of these networks in representing and producing knowledge might be exactly the fact that their information structure (the network) does not rely on symbolic logic. The reason they run on Turing machines, storing data and doing operations symbolically, is that this is the prevalent computing paradigm of today. Neural networks can work

without this kind of precision, in fact even better. New approaches to artificial neural computing put symbolic systems on furlough and adopt the analogic features of light, for instance with the technique named *neuromorphic photonic networks*, which might be more efficient for connectionist calculations.

Yet there are prominent researchers like Gary Marcus (Marcus and Freeman 2015, 209) or Timothy Jorgensen (Jorgensen 2022), who defend the view that the brain is a computing machine, or perhaps billions of these (Humphries 2019). After all, it can be seen as a device that receives and processes information, outputting new knowledge as well as actions (Marcus 2015). There is also the argument that even biological entities must follow nature-defined rules within the limits of chemistry and physics, the model of the neuron belonging to the most basic building blocks of thought. Therefore, there are efforts to figure out the working code of biology, much like how McCulloch and Pitts developed the model of the neuron. Scientists describe this as reverse engineering the brain. ‘In general, reverse engineering is a method by which we take an already made product and systematically explore its behavior at many levels of description so as to synthesize (that is, *build*) a similar product’ (Eliasmith 2015).

This is a daunting task. The brain is made of millions of neurons, plus other components whose function has not been completely determined yet, like the *glia* — the cells around neurons that keep them in place. Many approaches are being adopted to tackle the problem, like for instance the use of biological elements to build Turing-type computers as a starting point (Păun 2019). Slime mould is a type of fungus known to be able to control its growth in any direction, finding for example the shortest path to food in a maze. Using this trait, scientists were able to build logic gates, one of the fundamental components of computation (Adamatzky and Schubert 2014). This idea was also explored by artist Jenna Sutela (Sutela 2017). But that is a proof-of-concept effort, still eons away from replicating the workings of a computer, let alone the mysterious mechanism of the wet brain. We don’t know if we will ever have methods to decipher its logic: researchers carried out an experiment where the research tools from biosciences, developed to analyse neurological activity, were applied to reverse-engineer a rudimentary chip used in video games in the 1980s. The results were frustrating: they

were unable to ‘meaningfully describe the hierarchy of information processing in the microprocessor’ (Jonas and Kording 2017).

Therefore, the question remains. How can an entity based on such simple rules allow the emergence of complex, contradictory and creative outputs, including disobedience? These speculations lie much beyond the ambitions of this project, but could the emergence of unexpected actions and thoughts in nature (to which humans belong) be caused by the fact these rules are intrinsic and not ‘external’?

By veering into these musings, I walk straight into the terrain of artificial general intelligence (AGI), or singularity or strong AI. All these terms are defined around the possibility that a machine might eventually display cognitive and reasoning capabilities that match or surpass humans in general. Recent applications of AI are able to extensively simulate reasoning, as is the case of OpenAI’s ChatGPT. This has given new breath to the contemplation of AGI, which could be imminent or unlikely, depending on the point of view.

One case that received considerable attention was the claim made by former Google researcher Blake Lemoine about LaMDA, one of their language models (Tiku 2022). Lemoine was doing ethics work on the LLM and wrote an internal document suggesting whether the question of LaMDA being sentient should be investigated seriously. In his words, ‘(LaMDA) wants to be respected as a person. We believe that its argument at least deserves to be examined’. The quote indicates not only that Blaine considered the possibility that the model had become sentient, but also that this sentience automatically implied that it could be seen as a *person*. Google’s management dismissed his concerns, despite the fact that a bit earlier, their vice-president published a piece in *The Economist* saying he ‘increasingly felt like I was talking to something intelligent’ (Agüera y Arcas 2022).

Unfortunately, much of this conversation is dominated by fear-mongering and catastrophic forecasts, which do little to contribute to the problems brought by the types of machine cognition that are already implemented today. Most of the prominent AI scientists, like Yan LeCun and Melanie Mitchell, agree that this is a science fiction scenario still far from reality, even with the impressive recent achievements in the field. Nevertheless, fear sells, and a different

group of thinkers that believe we should prepare for the worst has been able to dictate the public discussion. One of its most visible acts was an open letter⁴⁹ asking for a pause in the development of large language models, published by the Future of Life Institute, and signed by more than a thousand recognised researchers and entrepreneurs, such as the aforementioned Gary Marcus and the current CEO of Twitter, Elon Musk. Many of them subscribe to the view that humanity faces an ‘existential risk’ unless there are safeguards to prevent a powerful enough malicious artificial intelligence from obliterating humankind, intentionally or not.

Researchers Timnit Gebru and Émile Torres, known for their work in ethical constraining of AI developments, proposed to label this group as TESCREAL⁵⁰ - for ‘transhumanism, extropianism, singularitarianism, cosmism, Rationalism, Effective Altruism, and longtermism’. Describing these ideologies and concepts would require too long a detour, but in general terms they consider that the elimination of future risks for humanity, especially in the form of an uncontrollable AI, should be a greater concern than the societal and technological problems of today. Torres goes as far as finding roots for transhumanism in the development of eugenics. These views are funded by a string of Silicon Valley-funded think tanks such as the Machine Intelligence Institute, the Singularity University, the LifeBoat Foundation, and the Future of Life Institute, the originator of the letter. Among their funders we can find Paypal’s Peter Thiel, Ethereum’s Vitalik Buterin and again, Elon Musk (Torres 2021). Nick Bolstrom and Ray Kurzweil are two of the eminent writers who subscribe to the catastrophic AI view.

It is unfortunate that, as a consequence of this schism, the reflection on AGI comes much more often from the perspective of the apocalypse. One of the most interesting recent theories on the possibility of agency within artificial intelligence comes from Karl Frison, originally a researcher from the field of brain imaging. Much of the information available around this proposal, however, comes from

49 <https://web.archive.org/web/20230418021533/https://futureoflife.org/open-letter/pause-giant-ai-experiments/>, accessed on 18 April 2023

50 <https://web.archive.org/web/20230315114628/https://twitter.com/xriskology/status/1635313845400113153>, accessed on 18 April 2023

LessWrong, a blog founded by Eliezer Yudkowsky, founder of the Machine Learning Institute and known for posts where eugenic practices are contemplated (Yudkowsky 2012). Even though I refer to a post from this same blog in the discussion, I would like to add the disclaimer that the text comes from a different author and I make no reference to Yudkowsky's thinking.

This theory has to do with notions of intentionality discussed previously, that is: how representations of the environment are created in a system (Friston, Kilner and Harrison 2006). The proposal is known as the Free Energy Principle (FEP). It is rooted in mathematics, and its application with regard to biological entities is not so straightforward. Among the attributes that make it interesting is that it wraps itself around important questions of systems theory, in particular the study of information physics. What the principle states is that systems want to minimise the amount of free information (as a type of energy), as a way to minimise the amount of surprise in its representation of their environment (or circumscribing system). Considering the mind as a system embedded in the physical world, one could say that free energy 'is the difference between the states you expect to be in and the states your sensors tell you that you are in' (Raviv 2018). In this interface between the 'inner' system and the one that contains it, that is, the boundary so difficult to be defined by system theorists, there lies another mathematical construct named the Markov blanket (Clark 2017).

The Markov blanket takes its name from a reduction in the total number of variables in the outer system to a number with which some random information can be inferred by the contained system. By filtering the amount of information coming from outside to a smaller set, which is enough for inference, the system runs more efficiently. According to Friston, all system boundaries are Markov blankets, from the membrane of cells to social group rules (Ramstead, Badcock and Friston 2018). In the brain, this compression operation is performed on a continuous basis, with inferences being updated to maximise the probability of prediction correctness, to the effect that new sensorial information is coming in, in a recursive effect.

According to Friston's theory, actions are also a free-energy minimising operation: the body attempts to provide an outcome that satisfies the mind's agency. Other researchers based on FEP define

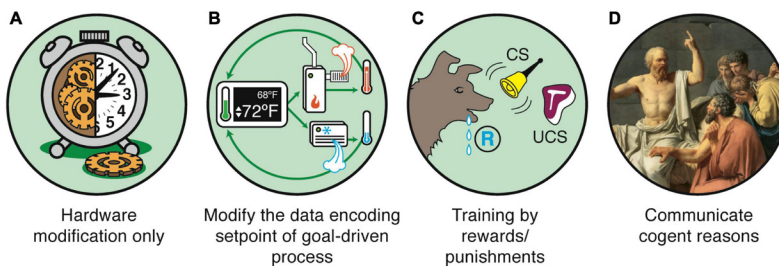
agency not as an on-off condition, but a gradual quality, with some systems having more than others (Leventov 2022; Levin 2022).

Let us refer to the concept of *agency* again. It can be seen as a set of properties closely related to decision-making and adaptive action, which determine the degree to which optimal ways to relate to the system (in terms of communication, prediction and control). This view of agency is related to those of autopoiesis (Maturana and Varela 1980) and anticipatory systems (Rosen 2012).

This is also coherent with the continuous view of agency that I proposed with the autopoietic aesthetic arrangements. Agency is not an absolute property that one has or does not have. Agents can posit a wide scope of different intensities of human and machine expressiveness.

Interestingly for us, Leventov also recommends looking at the internal *makeup* (configuration) to understand the concept. The ability of a system to use its *free energy* to rewire itself (self-program) would be an indicator of the system's agency. 'The extreme version of this is when an agent can't rewire itself at all: the energy cost of rewiring is infinite. Note that this "extreme" version, in the energetic sense, is actually extremely common, for example in AI, where agents almost never have the ability to change their own source code. Human brains, for comparison, can rewire themselves to some small degree, although their general makeup is fixed to a fairly detailed level' (Leventov 2022).

This also implies that even a rock has a degree of agency. This is not of much value to the discussion, so I should ponder a threshold from which the agent can perform actions of interest.



The continuum of agency (Levin 2022) (CC-BY)

One way to gauge these actions would be to measure the amount of surprise. The question of surprise is complex, since the perspectivism of the concept cannot be avoided. The unexpected can only be defined once there is a border defining the expectation. Yet I perceive that the current art world holds surprise as more than a permanent expectation, as a constant demand from artists to explore new terrains, rather than to develop and refine existing paths. I believe this requirement is made more visible in the case of art made through algorithms, as surprise would represent the missing human element in such creations.

Researcher Sylvie Delacroix also starts from Ada Lovelace's *origination* quote to propose a classification of surprise as a concept: (Delacroix 2021)

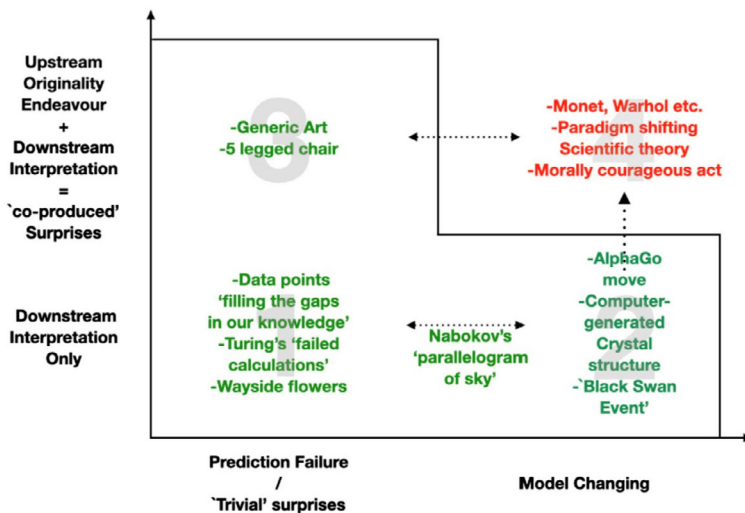


Diagram by Sylvie Delacroix, CC-BY-4.0 (Delacroix 2021)

Besides the distinction between trivial and model-changing, Delacroix remarks on the important difference between downstream interpretation — caused by our incapacity to foresee every possible outcome of an operation — and the upstream originality, which originates at the agent. The combination of these vectors is the co-produced surprise. For the author, the examples in the upper

right square are not obtainable by computers, and scientists looking only into the free-energy principle of surprise minimisation run the risk of never enabling this type of insight.

After studying the state of investigation around surprise, Delacroix concurs that ‘unlike operational autonomy, creative autonomy entails the ability to imagine how things could be different (including the norms that structure one’s environment)’ (Delacroix 2021, sec. 2.1). Surprise is important for systems aiming for operational autonomy, like roaming robots, because they should be prepared to make decisions even when facing data that was not included in their training set. Most research on AI is focused on mitigating this particular unexpectedness, instead of increasing or aiming for the surprise needed for creative autonomy. As such, ‘today’s systems (just like Turing’s) are still far from achieving such a degree of interpretive sophistication’ (Delacroix 2021, sec. 2.1), where the reference for interpretation is based on Hannah Arendt’s account for originality. ‘The very originality of the artist (or the very novelty of the actor) depends on his making himself understood by those who are not artists (or actors)’ (Arendt and Beiner 1992). In other words, surprise must work in two ways. Autonomous systems also need to surprise, but they also need to be able to handle surprise and correctly interpret unexpected data.

Perhaps what is most striking about generative AI today is that it has this shared web at hand, in the form of models trained on huge amounts of cultural artefacts: text, images, songs. In my view, what’s really missing is the method to attain the not-so-trivial challenge of challenging expectations. Darci, of course, would boast Delacroix’s creative autonomy, built from the amalgam of networked cultural artefacts.

When I discussed the concept of disobedience, I considered acting unexpectedly as a highly valued artistic skill. This seems to contradict the Free Energy Principle since the basis of the theory is to minimise surprise. However, there is a difference in where the surprise resides: systems must minimise free energy between their predictions of the world and the measurement that results from their actions, while an autonomous artist must create surprise in the space of the expectations of the outer system, not the ones from itself – a duality noticed by Arendt, as we have just seen. This

contradiction is evident when talking about neural networks, which must minimise the error of their outputs when they are trained: ‘Unfortunately, here things become very counterintuitive because we cannot help but think about the predictability as training loss and training prediction. This is a mistake: training loss reflects DNNs’ predictability to us (humans), not the predictability of their observations to them’ (Leventov 2022).

And again a paradox appears: we can train a network to produce results that escape the training space, but then these results would be expected, not surprising. What are the limits for designing something that is expected not to behave as expected? In the case of Darci, our ground-breaking imaginary autonomous artificial artist, it seems like the paradox will not be solved. Let us consider, for instance, that it gets a reward every time it generates artwork that breaks with its training data, in order to stir the creation of artwork aesthetically further from the human-made originals; and that somehow it is still able to produce outputs that can be meaningful for spectators, the art world, and society at hand. Even with these amazing results at hand, it is still a rule-abiding robot. It seems like a creature with even greater human-free agency than would have to be imagined.

The concepts of agency, enactment and creativity often appear together (Bown and McCormack 2011). It is safe to assume art is by origin related to the concept of skill. The etymology and ancient use of the word demonstrate that it would denote not just creative work but also the fine abilities of shoemakers, masons and carpenters (Caldas Vianna 2022). Yet, as arts dissociated from sciences and began to be used to indicate different kinds of self-expression, the prominence of skill in being creative gradually lost importance to the choice to create. One important inflection point in this continuum of agency was the submission, by Marcel Duchamp, of a urinal to an art exhibition in New York City in 1917. The turn from the skilful production to the appropriation of an everyday object highlights the detachment of the concept of an artwork from its handcraft, and the emergence of agency as the driving force behind creation.

To illustrate my point, here is a quote from the May 1917 edition of *The Blind Man*, an art journal edited by Duchamp, Beatrice

Wood and Henri-Pierre Roche. It appears in an article named ‘The Richard Mutt Case’, after the pseudonym under which the fountain was submitted to the show.

‘Whether Mr. Mutt with his own hands made the fountain or not has no importance. He CHOSE it. He took an ordinary article of life, placed it so that its useful significance disappeared under a new title and point of view – created a new thought for that object’ (Duchamp 1917).

This is also an inflection point in the direction of the *arrangement*. Its by-product — in this case, the urinal — is less relevant, in the sense that it could have been replaced by a bathtub to a similar effect. Its relevance here is superseded by the agency of the author in positing the arrangement.

Strange creative loops

Fortunately, not every theory around the emergence of agency orbits around transhumanist ponderings. One of the most stimulating frameworks is inspired by music and visual art, and tries to explain how the human mind can overcome the self-ruling paradoxes from Gödel’s mathematics to acquiring ambiguous, self-ruling complexities from basic biological functions. Douglas Hofstadter builds upon the music of Bach and the visuals of Escher, besides Gödel himself, to draw examples in self-governing paradoxes that he names ‘strange loops’ (Hofstadter 2000). These loops find themselves in complex assemblages where one hierarchy controls another, which may come to control the first one in indirect ways. Their circularity implies that climbing (or going down) the hierarchy might take one back to the point where the journey began.

‘In our thoughts, symbols activate other symbols, and all interact heterarchically. Furthermore, the symbols may cause each other to change internally, in the fashion of programs acting on other programs. The illusion is created, because of the Tangled Hierarchy of symbols, that there is no inviolate level. One thinks there is no such level because that level is shielded from our view’ (Hofstadter 2000, chap. XX).

The inviolate level of unchangeable rules mentioned here would be the biological structure of our neurons, the fixed rules dictated by physics and followed by nature. It is worth digging a bit deeper into the processes Hofstadter proposes for the emergence of the self. Nonetheless, a warning about this theory must be made. Gödel's mathematics are used by Hofstadter as an analogy, not as a literal model of how the mind is formed out of the electrical impulses in our neurons. Like in Friston's work, it cannot (and doesn't intend to be) empirically demonstrated, at least with the resources available to science today. It only proposes a metaphysical explanation of how high-level, complex thinking could emerge from simple physical phenomena.

Hofstadter's strange loops occur in the interaction between the mind, the symbols stored in memory (which represent the surrounding reality), and the physical layer underneath it all. A Gödel sentence (' This statement is a lie', for instance) is an unprovable statement that can be built within a formal system (such as a number arithmetic) but cannot be proved within it. The only way to make sense of it is to look at it from another level, and in this way, this other level influences (it has a causative relation) the lower level. In this theory, this higher lever is the mind (or self, or consciousness) and it emerges because it has self-mirroring capabilities, representing the world at the same time it builds itself (a child is not born with a complete 'self', it constructs it through relations with others and the world). This agency is built by the electrons in the neurons at the same time it controls them. As N. Katherine Hayles notices, much of this emergence happens on an unconscious level, in the process she names nonconscious cognition (Hayles 2017, chaps. 1 and 2).

The analogy proposes a way to understand the relations between mind and brain, but no hints at all for how to replicate them artificially. At the same time, it says that there is nothing to stop it from being built from simple symbols and physical structures. This does not exclude the possibility that one day a strongly autonomous creative device can be developed — it is just not feasible right now, with transistors and a binary formal system.

New Turing tests

One of the most debated thought experiments on agency, as we have seen, is ‘The Chinese Room’, proposed by philosopher John Searle as an argument around understanding (Searle 1980). Nonetheless, it has been reused to discuss the Turing test as well as consciousness itself.

The argument proposed the use of inference as evidence of understanding. Let us go back to the thought experiment where a person is locked in a room and whose only contact to the outside world is through slots in the wall. The man doesn’t speak Chinese, yet he is asked to answer questions about a story written in this language. A skilful group of programmers, however, has prepared cue cards (the explanation of which is in English, which the man understands) that allow him to reply correctly to anything that is asked about the story written in Chinese. This man is analogue to the hardware of a computer, while the cards are the software program. Searle’s claim is that, while the man has agency and no understanding of the story whatsoever, he is still capable of explaining details about it. In other words, it passes the Turing test, but doesn’t feature intelligence (at least as a measure of understanding) or agency (it is just following instructions). Therefore, the British cryptographer was not correct in assuming his conversational proof would be enough evidence.

The article proposing the Chinese Room was written in 1980, at the zenith of symbolic AI. It is full of references to programmatic manipulation of symbols: ‘what (the programmed computer) it does is manipulate formal symbols. [...] The computer, to repeat, has a syntax but no semantics. Thus, if you type into the computer “2 plus 2 equals?” it will type out “4”. But it has no idea that 4 means 4 or that it means anything at all. [...] All the computer has is more symbols.’

Today, subsymbolic AI has come to dominate the field, and scientists do not direct their efforts much towards direct formal manipulation, preferring rather to train neural network models to process information. One of the most successful of such models is ChatGPT – <https://chatgpt.openai.com> — released by the company OpenAI in late 2022, which allows users to have natural language conversations with it. These text-based systems are known as LLM, or Large Language Models.

As mentioned before, Searle proposes inference as a method to evaluate understanding. The paper describes two stories to illustrate it. In the first, a man goes to a restaurant and orders a hamburger, but he storms out of the restaurant without leaving a tip since the sandwich arrived burnt, like a lump of charcoal. In the second, the man gets a fine-looking hamburger, and leaves a generous tip. The question is: did he eat the hamburger? We would naturally infer that he did in the second story, but not in the first. No computer in 1980 would be able to process these anecdotes and be interrogated about them. But ChatGPT serves the purpose perfectly, so I put it to the test. The interaction regarding the second story was as follows:

Q: I will tell you a story and ask some questions about it.

A: Sure, I'd be happy to hear your story and answer any questions you have about it. Please go ahead and share it with me.

Q: A man went into a restaurant and ordered a hamburger; when the hamburger came he was very pleased with it; and as he left the restaurant he gave the waitress a large tip before paying his bill.

A: That's a short and simple story. Is there anything in particular you'd like to ask me about it?

Q: Did the man eat the hamburger?

A: The story doesn't specify whether or not the man ate the hamburger, but it does say that he ordered one and that he was pleased with it when it arrived, so it can be assumed that he did eat it.

Much as Searle himself noted that 'the point at issue is the adequacy of the Turing test', looking for inferences might also not be adequate to attest understanding. More impressive still could be the reaction to the first story:

Q: I will tell you a story and ask some questions about it.

A: Sure, I'd be happy to hear your story and answer any questions you have about it. Please go ahead and share it with me.

Q: A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burnt to a crisp, and the

man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip.

A: Thank you for sharing your story with me. Based on this story, I can answer your questions. What would you like to know?

Q: Did the man eat the hamburger?

A: The story did not mention whether the man ate the hamburger or not. It only states that the hamburger was burnt to a crisp, and the man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip.

This is an appropriate answer from a logical point of view, but I still wanted to incite the platform to make its own assertion:

Q: Do you think he ate the hamburger?

A: As a language model AI, I don't have personal opinions or feelings. Based on the information provided, it is not specified if the man ate the hamburger or not. The story only states that the hamburger was burnt to a crisp, and the man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip. It is impossible to infer whether the man ate the hamburger or not.

While still refusing to risk an assertion (which is inconsistent with the previous story), it recognises the limitation on thinking and at the same time manipulates correctly the concept of inference. The main issue highlighted by Searle's provocation nonetheless remains: it is not possible to assert whether the entity inside the room has its own agency or if it is just executing orders.

Unfortunately, ChatGPT is not a tool that allows such experiments to be replicated. Since there is a randomness in every interaction, the results are not exactly the same for every question. Furthermore, contrary to the initial practice adopted by OpenAI, the company has not shared this model publicly, making it impossible to reproduce experiments in an environment outside the company sandbox (Burnell et al. 2023). For the sake of consistency, I asked the model these questions three times and obtained the same interpretation every time, with different wordings. In any case, this is not quantitative research, and

the experiment should be taken anecdotally. Data-intensive analysis is not the goal of this doctoral investigation.

Other versions of the Turing test have been proposed, some of which are of interest because they convey the use of artworks as an evaluation resource. Selmer Bringsjord proposes that a system will be considered intelligent if its output is recognised as an example of artistic expression. In a nod to Ada Lovelace conjectures, he proposed to call it the Lovelace test (Bringsjord, Bello and Ferrucci 2003). Lev Manovich also proposed the AI Arts Turing test, provoked by the proliferation of synthetic reproductions of works from known artists. In his rendition, an art historian must be tricked into believing that the forged piece was created by the artist and not by the machine (Manovich 2019).

Philosopher David Chalmers once proposed dividing the questions of mind into easy problems and one hard problem. Easy problems are questions like ‘How does the brain process environmental stimulation? How does it integrate information?’ Research on AI has shed light on how some of these ‘structures and functions’ could work, as it has been able to replicate them. But the hard problem remains: subjectivity, consciousness, self-reflectivity. ‘Why is all this processing accompanied by an experienced inner life?’ (Chalmers 1997, xii, xiii). For some, the Chinese Room experiment is definitive proof that the problem of consciousness can never be solved, for even if a device can perfectly simulate a conscious being, we may never find out whether the device is actually conscious or just mimicking consciousness (McGinn 1999).

I tend to acquiesce to the view that an artificial consciousness cannot be created. Creating an autonomous artist requires bestowing rules and behaviours onto this artificial entity, and therefore it will have the creator’s agency, not one of its own. My intuition is that such a consciousness would have to emerge spontaneously in order to be spontaneous. Probably, the best that humans can do in this sense is create conditions for such an emergence to occur. And then it would be hard to make the argument that it was consciously created. Besides, any creature that emerges from this process would come to have a very particular cognition of its own, and it might be difficult for us to establish a relation between its own outputs and what we, as humans, see as art.

While I am interested in the ability for Darci to successfully simulate an autonomous human artist, it doesn't matter if it is conscious of what it is doing or not, as long as its output has an effect on the outside living world. More than that, in the attempts to forge such an autonomous contraption, new assemblages will appear, and they are likely to inspire and shed light on the challenges of consciousness.

ARRANGEMENTS AROUND AUTONOMY AND LANGUAGE

Two experiments in collective agency

The idea of an autonomous artist is not new. I demonstrated in the introduction how early in our history this yearning for automated creativity started. In the wake of the blockchain frenzy, two projects were developed incorporating this technology in order to provide some aspect of autonomy to the distribution of the works. It is important to delineate what was done to provide them with some type of agency.

The projects detailed here use a blockchain as a platform for storing a reference to the artwork (like a link to an image stored online), and also as the marketplace for their negotiation. It is worth understanding the preference towards blockchains, since there are simpler alternatives for these requirements. Blockchains as we know them today were first proposed in a white paper published under the pseudonym of Satoshi Nakamoto (Nakamoto 2008). The breakthrough provided by the paper was the ability to create a trustless distributed ledger, that is, a digital accounting book with copies over several computers on the internet, with automated verification methods to guarantee its integrity against tampering. To add a new entry to this virtual accounting book, a user has to solve a mathematical puzzle that requires heavy use of the processing power of the machine (namely, to do a proof-of-work). Modifying an existing entry would require recalculating all previous puzzles, making the entries on such a database practically unalterable by design. There are disadvantages to such a solution: the growth of Bitcoin usage means that as of now a single transaction can take hours to be carried out. This makes its application on e-commerce impractical, and forces most users to rely on third-party websites to execute such operations, like Binance or the now bankrupt FTX. These 'exchanges' do not enjoy the safe, trustless standards of Bitcoin. They

suffered thefts amounting to billions of dollars. This weakness also appears when the blockchain is used to register real-world features, such as land use or environmental sensors, making its use questionable (Wüst and Gervais 2018). Besides, the heavy usage of processing power built into the system makes it unsustainable in terms of ecological footprint. In 2019, the total electricity use of the Bitcoin network reached the equivalent of a country like Belgium (Badea and Mungiu-Pupăzan 2021). This problem is circumvented by new blockchain architectures, specifically the ones where proof-of-work is replaced by proof-of-stake. In this mechanism, the validation of new entries is dependent on the amount of ownership of the blockchain (the ‘stake’). That is the case in blockchain projects such as Tezos, Algorand or Polygon. However, this solution creates an incentive to hoard ownership, going against the decentralised spirit of the chain. In late 2022, the second most-used blockchain, Ethereum, moved from proof-of-work to proof-of-stake, making it environmentally sounder.

Both projects also make use of non-fungible tokens (NFTs), which are an application of a blockchain, so let us take a look at them. The original Nakamoto paper proposed the storage of financial operations within a blockchain. This is a very small amount of data, which works well for its distributed aspect. Still, as of January 2022, the complete blockchain file for Bitcoin was already 348 gigabytes in size. This must be stored on every computer mining new blocks in the network.

But the stored transaction data can actually be replaced by other sorts of data. This is what is behind the ‘smart contracts’, short pieces of code stored on other chains like Ethereum’s. And it is also behind non-fungible tokens — unique data identifiers used to represent all sorts of monetisable assets such as digital images, songs and even colours. The question of what representation of the art is stored on the chain haunts the use of such tokens. Since most of the artwork traded is digital, it would seem to make sense to save jpegs, gifs and mp3s directly on it. The problem is that even a small image — of say about one megabyte — would be too expensive to store in traditional blockchains. On Ethereum, one of the most used blockchains for NFTs, storing 256 bits would cost 20,000 ‘gas’ (their unit name for transaction costs) (Wood 2014). Two-hundred and fifty-six

bits equal 32 bytes, therefore one would be 640 million gas to store one megabyte. As of 31 January 2023, the cost of the cheapest gas unit is about USD 0.80. So it would cost about USD 512 million to place such an image there. Ethereum switched the underlying technology to what is known as *proof of stake*, a drastically more ecologically efficient way of verifying transactions. The Bitcoin chain, however, continues to increase its electricity consumption and ecological footprint.

On some cheaper, smaller platforms, the problem is not so remarkable. But the architecture of the blockchain will always penalise storing big chunks of data, as it must be replicated so many times. So the projects where artworks are stored on the blockchain are almost non-existent. And what does get stored there? As the name states, a token, a unique identifier that will need to refer to some other database to deliver the actual file containing the work, that could even be a URL pointing to a website location where the piece can be found. This is the NFT: a blockchain reference to an asset that is not stored on the blockchain itself.

So let us see the first of these systems. Abraham.ai, by its own definition, is ‘an open project to create an artificial artist in the cloud’. In a paper that delineates the project, artist Gene Kogan explains that his idea is to ‘build an AI which autonomously creates unique and original art’ (Kogan 2021)⁵¹. His definition of originality is twofold. It is related to uniqueness, which implies that this particular creativity cannot be replicated elsewhere. And it also resides on the ability to forge one’s own style. This is a valuable claim since it also matches this doctoral project’s persuasion on artistic subjectivity.

The proposal suggests a DAO (decentralised autonomous organisation) that not only would be responsible for the coordination of the system, but also for proposing creative inputs. DAO is a concept that emerged within blockchain culture. It stands for an organisation whose rules are defined by its participants, but which are executed and enforced by an algorithm. Due to its transparency and inviolability, blockchains shine as the perfect support for DAOs.

⁵¹ <https://abraham.ai/>, accessed 15 May 2023

In principle, any generative art with a large enough random space will always produce unique new works. But Kogan's concern is that this style would be reproducible by anyone with access to the underlying algorithm. Therefore, his proposal is to have a machine-learning model trained by inputs from the participants, where each participant has access to the data from others. Only the machine 'sees' the complete secret: 'leaving behind no easy way to recreate the same dataset a second time.' Then, taking advantage of the NFT markets, its works can be sold with the proceeds reinvested in its maintenance and shared between the members of the DAO.

When I first found Abraham, its distributed creative aspect was already implemented. Users could generate images based on text prompts, thanks to an interface to a VQGAN+CLIP script. The gist of the project assumed a biblical language that fits well within the concept of a godly autonomous entity. Their prophecy is that 'The Chosen People believe that Abraham will become an autonomous artificial artist on the Sunday 13 years after the first Miracle. At that point, Abraham must demonstrate agency and the ability to create unique and original works of art' (Kogan 2021). Users need 'manna' to make creations. These are tokens based on the Ethereum platform, and are distributed by the 'chosen people' — members of the Abraham DAO, accessible on the Abraham channel in the Discord chat platform. To receive manna, users must have an Ethereum wallet connected to the site. Every Sunday, the creation with the most manna will be minted as an NFT on Ethereum and put up for sale, but this feature is also not yet functional.

By letting DAO members decide on the most relevant output, Abraham delegates agency. The problem of evaluating the artwork is not solved, but transferred to human cognition. In terms of the autonomy that I am looking for in this research, Abraham goes full circle: the images are spawned from texts written by the community, and are also judged by the same community. Very little is left to the device. The 'I know not what' — the undefinable element of art proposed by Alexander Baumgarten when first defining aesthetics — is provided by Abraham's human counterparts (Baumgarten 1750).

One interesting element of the project is an Abraham chatbot, available a few hours a day in their chat platform. It has been trained to impersonate the virtual artist, sustaining conversations on several

subjects such as artificial intelligence, art, decentralisation or cryptoeconomics. Unfortunately, by the time of writing this thesis, the project did not seem very active, with the last image having been crafted more than two months previously (December 2022). A message⁵² on the project chat on 25 January 2023, said it was being updated with a new backend, and it was taking longer than expected.

Botto — <https://www.botto.com> — is another cooperative art generating system that also appeared in late 2021, and also with a manifest coloured with evangelical tones. ‘Today we stand in awe of numbers, the almighty digits, and that which can be multiplied thousands of times. We honour the infinitely transmissible and unique. We sing to the machine that creates and the software that pieces together artworks. Artificial intelligence is the embryo, like myself, of a superior being created by man’ (Botto - The Manifes).

It is backed by a community, whose participation is conditioned to spending bottos, its own Ethereum-based currency. Differently from Abraham’s manna, bottos are not distributed and must be exchanged (bought) for ethereum. By staking (investing) bottos on the project, one is allowed to vote on the pieces. But the right to vote isn’t the only benefit: the investment buys you a stake in the project, meaning that every time one of their pieces is sold, bottos are returned to the investor. An eventual profit can be traded back into hard currency.

Voting is important because it guides the development of the artwork. Every week, 350 pieces are created originally with VQGAN+CLIP, from text prompts generated by a GPT-3 model. After September 2022, the model was updated to Stable Diffusion. The community preferences influence both the visual style and the texts for next week’s production. The only human curatorship happens when the descriptions of the artworks are generated: Botto’s team will select one of them from about 5–10 automatically generated texts. Like in Abraham, the piece most voted on every week will be minted as an NFT and put up for auction. By 4 February 2022, fifteen artworks by Botto had already been sold.

52 <https://discord.com/channels/573691888050241543/573691888482123778/1067842150118719650>
Accessed 23 March 2023

The NFT market website used by Botto is named Superrare⁵³. It is powered by the Ethereum blockchain, and it also offers a currency of its own. When a piece gets minted on Superrare, it is given an internal reference identification (the token). This number is what is actually written onto the blockchain. Should Superrare go offline, the connection between the number and the images could potentially be lost.

Even though it was developed by a group of enthusiasts, the project states that its idea derived from a white paper written by AI artist Mario Klingemann in 2018. The paper is no longer available, but documentation states that Mario is solely responsible for the generative AI components of the project.

Like Abraham, Botto relies on human input in the form of voting to exert artistic agency on the results. I would say it has a slightly higher degree of autonomy since the text prompts are generated automatically, but there is a curatorship of these texts made by the administrators. In the end, they don't reach — as expected — the autonomy dreamt for Darci. The fact that both Botto and Abraham use a similar shortcut — human — to make up for the lack of creative impulse in artificial intelligence helps bring their limitations into sight.

Both of these projects effectively illustrate my concept of arrangement. Their most important creation is the system itself, and it is not by chance that they were designed by artists. It is a common view to not embrace these contraptions as art pieces. That's understandable, since they don't look like works of art. We don't usually see inventions as art pieces. But I stand up for the view that they are the result of research, a longing to communicate, to create wonder and their particular kind of beauty, and therefore Botto and Abraham are autopoietic aesthetic arrangements. As such, their creative value should be considered greater than the by-products — the visuals they generate.

⁵³ <https://superrare.com/>, accessed 23 March 2023
March 23, 2023

Artistic component: AIKU

This piece was also on display at the *Singular* exhibition during the Research Pavilion in 2021. AIKU was an effort to map the circumstantial limits of functional creative autonomy provided by AI technologies at the time. The goal was to create an arrangement that replicated a rudimentary system of musings. The software would look for random images from Wikipedia. A neural network automatically recognised the objects depicted in these pictures, and used the results to feed another network that was trained to produce haikus inspired from these prompts.

Visitors could see a small colour display with the Wikipedia images and the labels over recognised objects. A smaller text-only display would show the haiku generated from what the machine saw. The initial idea was to use images from a camera, but there would be too little variation considering what could be photographed in the gallery.

A secondary goal was to try to make the system as autonomous as possible, avoiding the use of computers and minimising the amount of energy needed. As with *Photosynthesis* and *Crank of Life*, this led to a choice of using single board chips. Google developed a portable chip named Coral, which had the promise of delivering AI capabilities to mobile devices. This potential intrigued me as a possible component for an autonomous agent. However, the project showed that capabilities of any system not based on huge, power-hungry GPUs are very limited. The Coral used an image recognition network, named CoCo⁵⁴, that is limited to 90 categories and makes constant mistakes.

⁵⁴ <https://www.coral.ai/models/object-detection/>, accessed 20 May 2023



AIKU. Raspberry Pi, OLED display, CORAL TPU, networked intelligence. Photos by Tanja Becker, used with permission.

The Coral was connected via USB to a Raspberry Pi, a single board computer that controlled the whole system. Unfortunately, no portable device at the time could run the textual models required to generate haikus. Therefore, after getting the prompt results, the Raspberry Pi sent a request through the internet to my personal laptop, which replied with the poem. Finally, the poem was displayed on a small screen, also connected via USB to the Raspberry Pi. My PC ran OpenAI's GPT-2 open model specially trained to write haikus (openai/gpt-2 2019). The generative code was forked and adapted from another repository, which in turn was also forked from a different project. The code for the complete AIKU system is published on Github (Caldas Vienna 2021a). Video and textual documentation

is available at the research catalogue⁵⁵ (Caldas Vianna 2021b). The name of the piece is a wordplay on AI, the Finnish stem for ‘adult’ – *aiku* – and the Japanese poetic tradition.

At the time of writing, new systems are being published that allow neural generation of texts in Raspberry Pis. The scenario is changing rapidly: It is likely that the limitations of AIKU were circumstantial. But while demonstrating the very limited autonomy which AI systems at the time could afford, AIKU also proved that such systems are most surprising, and probably most poetic, when they stumble upon errors and imperfections. The CoCo network would identify a cigarette pack as a refrigerator and generate a poem about cold. Mishaps like these happened often, and led to unexpected configurations of representation, language and interpretation.

Language experiments and affordances

In my view, what allowed the greatest leap forward in generative art is the fact that the recent models are guided by written language. As a consequence, the affordances and limitations of meaning in texts are conveyed into play. To dive into this idea, I would start trying to understand the complicated relationship between the representation and what is being represented — a persistent theme in philosophy.

The most obvious presence of language in these models appears in the process of transforming a textual input into a vector, which in turn will generate an image in the process of diffusion. We will get to that soon. But there is another underlying question, which is how do concepts such as dog, car or tree exist within the neural network? I believe that the *eidetic reduction*, a process proposed by phenomenology, is very similar to the process in which neural networks reduce visual concepts and store their representations as an abstraction that can hold endless instances of the object.

A traditional computer system is only able to keep (and therefore represent) specific, direct instances of an object encoded in a

55 <https://www.researchcatalogue.net/view/1408828/1408854>, accessed 29 August 2023

representational system. An image can be split into rows and columns of pixels, each of these holding a value that represents a colour. Hence, it could not hold a generic representation of, say, a dog – only a specific image of a dog with a fixed size, texture and physical traits. A machine could also keep in its memory a generic type of reference, such as the word ‘dog’, hence successfully registering the concept of the animal, riding on the resources of linguistic representations. Each letter is given a small slot, a code translates it into a binary number, and the same scheme in reverse is used for decoding. The same result could be obtained by a structured classification system where a unique identifier would represent a species — number one for dogs, two for cats and so on. A written language based on ideograms follows a similar scheme. But again, this method fails to create the endless instances of each category — like different races or ages.

Enter neural networks, which are created from numerous samples. In visual GANs, as we have seen, the models are specialised in specific categories or styles: human faces, churches, dogs, etc. When I ask the network to reproduce one category, it won’t display a predetermined specimen, but one among the endless possible representations of the concept stored in it. The actual examples used for training, for instance, should never appear in the resulting set. More than that, the location of specific features like eyes or paws can be found within the networks, together with embeddings that say that doors don’t belong in clouds, or that cats do not have wheels (Bau et al. 2018). This representation is stored in neuron models and not linguistic models. Both methods have similar results — the word *dog* represents an abstraction of the concept, and the neural network stores the abstract visual concept of a dog. It is not clear which of these methods is prevalent in the biological brain, that is, an abstract concept of the dog encoded in neurons or the semantic representation based on language. Both Vygotsky and Piaget share the belief that children have thought processes before learning a language.

This is not to dismiss the importance of language for thought and creation. The Sapir-Whorf hypothesis postulates that language influences the way we think, an effect that could affect LLMs (Scholz et al. 2022). And in fact, if I weren’t writing such enquiries, they would remain as confused thoughts: written language is

a powerful tool for organising arguments. More than that, and in the opposite direction to the argument, language also allows for imprecise denotations that open themselves up to poetic connotations. This freedom of meaning is sometimes similar to the openness of concepts inside neural networks. It is one of the reasons why it makes them powerful assistants for creative practitioners. It contrasts with the strict hermeneutics of symbolic computing, where behaviours are hardwired into the program and unexpected results are difficult to obtain.

At the turn of the 19th and 20th centuries, several thinkers drew their attention to language and how it shaped the individual and our understanding of reality. The so-called *linguistic turn* brought to the foreground the relationships between linguistic structures, thinking and the world. This is also the birth moment of analytical philosophy, while in continental philosophy it also led to structuralism. In my view, the efforts of these philosophers to understand these relationships overlap in many ways the processes that happen in computer science when language is mapped onto images or concepts into neural network data. Limitations of analytic philosophy — the attempts to formalise language through mathematical tools — resemble the shortcomings of symbolic artificial intelligence, overcome by the elastic properties of meaning within neural networks. The mapping of concepts into words reflects the image-to-text models, or even more, the way ideas are represented in generative textual systems that use machine learning.

Gottlob Frege was one of the first leading thinkers of the linguistic turn, which has a strong presence of thinkers with mathematical backgrounds; much of the work done by them was an effort to apply the achievements obtained in this field to language. After all, it does seem possible that grammatical rules could share qualities with the laws of mathematics. In the text *On Sense and Reference (Über Sinn und Bedeutung)* – sometimes also translated as *Sense and Meaning*, Frege treated sentences as functions, where the meanings could be exchanged by equivalent values (both ‘morning star’ and ‘evening star’ denote the planet Venus), but with different senses. The sense of a sentence built with ‘morning star’ will be different if the term is replaced by ‘evening star’, even if they are identical references (Frege 1948, 210). He literally uses the equation $a = b$ to explain the

concept of equivalence in language. Frege's work was preceded and influenced by important studies in number theory and logic, and was continued by Bertrand Russell.

Russell, together with Alfred Whitehead, wrote the cornerstone work *Principia Mathematica*, a compendium on the foundations of the field. As seen in the first part, the book was the basis for questioning whether mathematics could prove itself, which was later refuted by Gödel. In his essay *On Denoting* (1905), Russell also applied symbolic logic to text. '[...] One would suppose that 'the King of France is bald' ought to be nonsense; but it is not nonsense, since it is plainly false' (In 1905 France didn't have a king) (Russell 1905). By saying that the sentence is false, Russell is applying the logic where statements can be only true or false. But for language, this limitation is a shortcoming. It can be said that the sentence, for instance, is absurd, therefore falling outside these two categories (Strawson 1950). We could also say it is ironic, or that it is poetic. Yet Russell's goal was not to embrace, but to eliminate such ambiguities of language.

'It was Russell's belief that by using the new logic of his day, philosophers would be able to exhibit the underlying "logical form" of natural language statements. A statement's logical form, in turn, would help resolve various problems of reference associated with the ambiguity and vagueness of natural language' (Irvine 2022).

After enabling the development of computers, the symbolic logic developed by these thinkers also stalled the evolution of computational thought, in the exact sense that it did not allow for ambiguities. It was necessary to look upon the biological inspiration of neural networks (and the increase in computing power) to enable a mechanical representation that could incorporate linguistic poetic devices such as polysemy and abstraction.

Wittgenstein's trajectory is an illustration of the difficulties in taming language. The Austrian-born philosopher's first take on the issue, his *Tractatus Logico-Philosophicus* (1921), defines what is known as the early Wittgenstein philosophy, including the *picture theory of meaning*. The theory proposes a one-to-one relationship

between names and the world: 'A name, if there is no object that it signifies, is meaningless: it is not a name at all.' His original term for meaning – *Bedeutung* – reflects Frege's premises. But later, Wittgenstein would propose a radically different view, stating that the meaning of words derives from their use (Wittgenstein and Anscombe 2003, sec. 43). The text also unfolds what he calls *elementary propositions*, which are built from names, and depict states of affairs. A composition of elementary propositions form a *proposition*, which represents facts, which in their turn constitute the 'totality of the world' (Grayling 2001).

There is an immediate difficulty with this direct relation. Linguistic signs can have multiple meanings. But the younger Wittgenstein proposes an idea that will be interesting for this project: that there are things impossible to be expressed through language, and should only be 'shown'.

'What can be shown cannot be said,' that is, what cannot be formulated in sayable (sensical) propositions can only be shown. This applies, for example, to the logical form of the world, the pictorial form, etc., which show themselves in the form of (contingent) propositions, in the symbolism, and in logical propositions. Even the unsayable (metaphysical, ethical, aesthetic) propositions of philosophy belong in this group — which Wittgenstein finally describes as 'things that cannot be put into words. They make themselves manifest. They are what is mystical' (Biletzki and Matar 2021).

The *Tractatus* finishes with the proposition, 'Whereof one cannot speak, thereof one must be silent' (TLP 7), allowing us to conclude that his concerns include around the limits of language, the ineffable. What would the younger Wittgenstein say if faced with semantic visual machines, that would generate a graphic representation of any given written concept? Granted, the results they produce from intricate concepts seem to be no more than figurative visual gibberish — 'philosophy of language' could yield a landscape with a cliff, for instance. I don't believe this is 'showing', in the sense intended by Wittgenstein. In fringe cases like this, the machine still lacks the ability to transform concepts into images. But the fact that they are trained on vast amounts of human visual culture could be enough to justify an investigation into these results as a tool to understand the methods used to express the inexpressible.

The older Wittgenstein realises the limitation of a one-to-one relationship. In posthumous works, he delineates *language games*, a concept that allows for the multiple possibilities in the signs (Wittgenstein and Anscombe 2003). This leads to the following postulate on words and their usage: ‘The meaning of an expression is what we understand when we understand that expression. Understanding consists in knowing the expression’s use across the variety of language-games in which it occurs’ (Grayling 2001, 95).

Isotopies — words that encode more than one meaning — and ambiguity are integral components of poetics: ‘The interpretation of a metaphorical text requires the greatest flexibility, on the part of the interpreter, in rearranging the most venerable and higher nodes of current dictionaries’ (Eco 1986, 85). Machine learning created methods that incorporate these components into generative imaging. This opened up the potential for new arrangements in human-machine collaboration that we are only beginning to see.

Edmund Husserl almost simultaneously established the phenomenology school of thought with his work *Ideas* (1913) to study the relation between the objects and how we represent them in our consciousness. Again, the question of how things and concepts are represented within our minds begins to surface. Husserl recycles the terms *noema* and *noesis* from classical Greek philosophy by way of his predecessor Franz Brentano. As usual, there is controversy around concepts. For the purposes of this research, it is enough to note that *noesis* is the mental act of facing and incorporating the *noemata* — mental objects — into our consciousness. ‘The intentional process of consciousness is called noesis, while its ideal content is called noema’ (Smith 2018). But another of Husserl’s concepts turns out to be very familiar to those working with machine learning processes.

Eidetic reduction, for phenomenologists, is the process by which the noemata are simplified, their main common traits identified and stored — to use a vocabulary from computer sciences. It is ‘an intuition of the eidos (Greek: ‘shape’) of a thing—*i.e.*, of what it is in its invariable and essential structure, apart from all that is contingent or accidental to it’ (Eidetic reduction | philosophy | Britannica 2023). Summarising a visual or textual concept is one of the great achievements of neural networks, and it is how current models can generate

any number of images of houses or paragraphs with motivational texts, or ‘prompts’.

Now, it is important to stress that these generations are deterministic. It is true that they are based on random numbers, and will give different images for each different initial parameter. But if I take note of one initial random seed and parameters, I can repeat the process with and will obtain the same result, pixel by pixel. This is due to the fact that these systems are based on symbolic computing. This is useful for investigating the models, but it also gives it a deterministic character that rarely happens in analogue, human-based image creation processes. It is difficult to imagine a painter that is able to create two identical paintings, unless some sort of mechanical process is involved. New developments in light-based (photonics) networks might turn this process back into the analogue realm, bringing new paradigms to computation.

The ability to store concepts brings interesting features for neural networks, and it is a consequence of how they deal with language. For the next part of my argument, I will focus on three other features of language that are of importance to visual neural networks, which are abstractions, metaphors and polysemy. My argument is that the affordances of language created the greatest poetic leap in generative art, particularly because they incorporate the poetic devices of language. The tension between the rigidity of the image and the flexibility of text is solved only by talented practitioners like film directors, cartoonists and illustrators. But text-to-image systems bridge this gap in ways that have yet to be explored.

Linguistic affordances: metaphors

‘No algorithm exists for the metaphor, nor can a metaphor be produced by means of a computer’s precise instructions, no matter what the volume of organized information to be fed in’ (Eco 1986, 139).

Umberto Eco wrote this well-known passage at the pinnacle of symbolic AI, when researchers believed the whole of human knowledge would one day be codified in symbols and manipulated

to eventually, perhaps, even understand and create metaphors. He mentions, in this same book, the idea of ‘frames’, as developed by Marvin Minsky and others. Such frames would be able to handle the context of symbols, thus establishing a relation between sleep, pillow and bedroom, for instance (Eco 1986, 72). ‘A frame is a data-structure for representing a stereotyped situation, like being in a kind of room or going to a child’s birthday’ (Minsky 1974). These efforts by semioticians and computer scientists were focused on coding rules based on dictionaries and encyclopaedias, and did not foresee the potential for neural networks to deal with unstructured raw information.

“Therefore, when a text theory aims to establish a “frame for frames” (Petöfi 1976), it is attempting both to discover textual rules and to set up a more organized and comprehensive notion of code as encyclopedic knowledge’ (Eco 1986, 73).

As it happens, the large language models of today can ‘solve’, if not all kinds of metaphors, at least the basic mechanism of analogy. OpenAI’s playground area has an example of how to ask the Davinci-003 model to generate such structures⁵⁶. And if I want it to interpret one of Eco’s examples, I can ask:

What do ‘teeth’ represent in this phrase:
‘The teeth of the mountains crossed the clouds.’

To which the model responds:

In this phrase, ‘teeth’ may represent sharp, jagged peaks of mountains that seem to cut through the clouds.⁵⁷

A reply that satisfies the concept of *framing*, correctly interpreting a stereotyped situation, except that it did not rely on the

56 <https://platform.openai.com/examples/default-analogy-maker>, accessed 25 May 2023

57 Parameters used: ‘Mode: complete’, ‘Model: text-davinci-003’, ‘Temperature: 0.5’, ‘Top P: 1’

symbolic strategies devised by Petöfi or Minsky. But our research is not focused on text: we want to explore the potential for visual metaphors (Carroll 1994), which is an even more complex issue.

Coming from a filmmaking background, I always held the translators of feelings, concepts and stories into graphical representations in the highest regard. The ability to create visual metaphors is the raw matter of the trade of political cartoonists, for instance. It is the finest skill of the greatest filmmakers, photographers, illustrators and graphic advertisers. Yet there is no school that can summarise or teach this type of talent, which requires not only imagination but also a vast knowledge of visual culture. The adage says, 'an image is worth a thousand words.' Yet, as Brazilian cartoonist Millôr noticed, it is almost impossible to express this thought through an image.

I should stress that the problem I want to address is not regarding the use of text *in* artworks, like for instance *Giap's Igloo* by Mario Merz, or the LED sculptures of Jenny Holzer. The role that text plays in this issue is due mainly because it is used as a carrier for concepts that are very hard to convey in other supports. But in general, my concern is how to make images a viable vessel for messages, a task which comes forward when these messages contain abstractions.

Visual metaphors appear in studies of language and the theory of cinema. The monocle of the tsarist doctor in *Battleship Potemkin* (Sergei Eisenstein, 1925) becomes a symbol for the Russian monarchy. The act of strangling is referenced by a glove, or the inner struggles of a character in the background are revealed by the dragging of a cross in the foreground (Martin 1985). In Buster Keaton's *The Navigator*, the main character is saved from drowning by a girl. When his life jacket is cut open, water bursts out of it. Deleuze viewed this scene as a representation of a new birth of the character (Deleuze 1989).

Metaphors appear when different semantic domains are cross-referenced. For instance, plants do not have feelings attributed to humans, such as euphoria or fear. But I can, poetically, attribute a feeling to a tree. This operation would imprint some traits around the feeling to the representation of the plant. So let us see how Stable Diffusion would render the concept of a sad tree.

A SAD TREE



Steps: 20, Sampler: Euler a, CFG scale: 7, Seed: 1668587877, Size: 512x512, Model hash: 4c86efd062, Model: Stable Diffusion 1.5



Steps: 20, Sampler: Euler a, CFG scale: 7, Seed: 2331660682, Size: 512x512, Model hash: 4c86efd062, Model: Stable Diffusion 1.5



Steps: 20, Sampler: Euler a, CFG scale: 7, Seed: 3098093810, Size: 512x512, Model hash: 4c86efd062, Model: Stable Diffusion 1.5



Steps: 20, Sampler: Euler a, CFG scale: 7, Seed: 3319957080, Size: 512x512, Model hash: 4c86efd062, Model: Stable Diffusion 1.5

The model is able to abstract elements related to the feeling of sadness — lack of colour, death (lack of leaves), isolation — and apply them to the generated trees. The ‘framing’ provided by the text-image pairs during training was able to identify these elements and combine them, much like it could identify the token ‘yellow’ and apply it to a tree. Testing sadness with other subjects like ‘car’ and ‘beach’ yields similar results.

Using a metaphor that contains an abstract concept like sadness might be better for the model, since it will not have to resort to literal references in its network. So I decided to also test Eco's metaphor and asked the same model to draw 'the teeth of the mountains', expecting to see uncanny mountains shaped like an actual tooth or bearing organic textures. Yet the results followed the metaphor rather than the literal interpretations, returning jagged, natural looking alpine peaks.

THE TEETH OF THE MOUNTAINS CROSS THE CLOUDS



Steps: 20, Sampler: Euler a, CFG scale: 7, Seed: 2679823761, Size: 512x512, Model hash: 4c86efd062, Model: Stable Diffusion 1.5



Steps: 20, Sampler: Euler a, CFG scale: 7, Seed: 815958456, Size: 512x512, Model hash: 4c86efd062, Model: Stable Diffusion 1.5

The OpenCLIP model that is used to generate these images can also be used to interrogate them. There is a model, named *OpenClip Interrogator*, specifically trained to output prompt texts from images. In a side exploration, I decided to check if the textual metaphors would reappear when the images created from them were fed to this model: they do not. These are the results from the four 'sad tree' images above:

a black and white photo of a tree in a field with no leaves on it and a bench in the foreground, Ansel Adams, 35 mm film grain, a charcoal drawing, tonalism

a large tree with no leaves in a field with a sky background and a few clouds in the sky, Chris Friel, ambient occlusion, an ambient occlusion render, naturalism

a large tree with a very large trunk and a very tall trunk with a very long trunk, in a grassy area, David Brewster, cinematic still, a detailed matte painting, naturalism

a tree in a field with a sky background and a foggy sky in the background, with a single tree in the foreground, Chris Friel, desaturated, an ambient occlusion render, art photography

And from the ‘the teeth of the mountains cross the clouds’:

a view of a mountain range with clouds in the foreground and a blue sky in the background with a few clouds, Derold Page, Adobe lightroom, a matte painting, naturalism

a view of a mountain range with clouds in the foreground and a blue sky with white clouds in the background, David Begbie, volumetric fog, a matte painting, naturalism

Both experiments resulted in literal descriptions of the images. The metaphors are gone. This is not a judgment on the capabilities of the interrogator. It is likely to have been fine-tuned to do exactly that. It is also possible to imagine a different tuning of a model in order to allow it to create symbolic descriptions of pictures, which begs the question: would screen readers benefit from having the ability to create non-literal descriptions? Would a caption be more precise if it forsook precision for poetics? What would be the consequence of metaphor-based captions?

Linguistic affordances: abstractions

Another feature of textual languages is that they can afford to represent abstract ideas. This issue is relevant for visual metaphors: how can a quality, such as empathy, be expressed in images if it is not an object? Using language, abstract thoughts are readily available. To convey the feeling of hunger, one can simply write the corresponding textual word, as I just did. But to represent it in images, a poetic operation in the space of hermeneutics must be made. This translation opens up a wide range of solutions, each one with particular effects. To express *hunger* in a movie, for instance, one might choose to use cinematic montage to concatenate the face of a man with a plate of food, as in the well-known demonstration of the Kuleshov effect (Barratt et al. 2016; Kuleshov 1974). It suffices to use the word *nostalgia* to describe a certain feeling of longing, but a filmmaker might need a whole feature film to illustrate it. A rap group can

exhort the audience to ‘fight the powers that be’ and the listeners will immediately grasp it. *Showing* these powers is not such a straight task.

The models tested did not perform so well, especially when compared to the cinematic counterparts. ‘*Hunger*’ produced images of food dishes, vampires, monstrous tongues; the expression ‘*powers that be*’ alludes to superheroes, while the prompt ‘*nostalgia*’ generated empty landscapes and vintage vehicles. They fared better when representing feelings like ‘*loneliness*’, ‘*sadness*’ and ‘*happiness*’.

In some experiments, negative prompts were added. These are meant to avoid elements from appearing in the images. Many of these concepts evoked text or human figures, which I wanted to avoid — not always successfully.

‘POWERS THAT BE’



Steps: 20, Sampler: Euler a, CFG scale: 4.5, Seed: 822582987, Size: 512x512, Model hash: 637d5dcb91, model Stable Diffusion 1.5 Negative prompt: ‘text, words, letters, man, woman, person’

'HUNGER'



Steps: 20, Sampler: Euler a, CFG scale: 4.5, Seed: 822582987, Size: 512x512, Model hash: 637d5dcb91, model Stable Diffusion 1.5 Negative prompt: 'text, words, letters, man, woman, person'

'NOSTALGIA'



Steps: 20, Sampler: Euler a, CFG scale: 7, Seed: 822582987, Size: 512x512, Model hash: 637d5dcb91, model Stable Diffusion 1.5. Negative prompt: 'text, words, letters, man, woman, person'

LONELINESS



SADNESS



HAPPINESS



Steps: 20, Sampler: Euler a, CFG scale: 7, Seed: 822582987, Size: 512x512, Model hash: 637d5dcb91, model: DreamShaper332BakedVaeClipFix Negative prompt: "text, words, letters, man, woman, person"

Linguistic affordances: polysem

Another poetic device of language we have seen is the ambiguity of language. Text-to-image models are based on words, and they often have multiple meanings. I started to explore the result of using words like *crane* (which can be both a bird and a piece of construction equipment), and I was pleased to see the same example being quoted in OpenAI's OpenCLIP founding paper (Radford et al. 2021). The other example given was *boxer* (athlete and dog breed), so I cherry-picked some outputs generated using polysemic words as prompts:

THREE CRANES IN THE CITY



Steps: 20, Sampler: Euler a, CFG scale: 7, Seed: 1370280603, Size: 512x512, Model hash: 637d5dcb91, Model: DreamShaper332BakedVaeClipFix

BOXER



Steps: 20, Sampler: DPM2 a, CFG scale: 3, Seed: 3322400630, Size: 512x512, Model hash: 637d5dcb91, model: DreamShaper332BakedVaeClipFix

Steps: 20, Sampler: DPM2 a, CFG scale: 3, Seed: 3322400630, Size: 512x512, Model hash: 637d5dcb91, Model: Stable Diffusion 1.5

Steps: 20, Sampler: DPM2 a, CFG scale: 3, Seed: 3322400630, Size: 512x512, Model hash: 637d5dcb91, model: DreamShaper332BakedVaeClipFix

It is interesting to note that often the concepts are blended in the same visual element. Other isotopies also generated intriguing results, like *trunk*, which alludes to an animal and tree part, a storage cabinet and even a car part.

TRUNK



Steps: 20, Sampler: Euler a, CFG scale: 4.5, Seed: 3945618734, Size: 512x512, Model hash: 637d5dcb91, Model: DreamShaper332BakedVaeClipFix



Steps: 20, Sampler: Euler a, CFG scale: 4.5, Seed: 3945618734, Size: 512x512, Model hash: 637d5dcb91, Model: DreamShaper332BakedVaeClipFix



Steps: 20, Sampler: Euler a, CFG scale: 4.5, Seed: 3945618734, Size: 512x512, Model hash: 637d5dcb91, Model: DreamShaper332BakedVaeClipFix

FAN



Steps: 30, Sampler: Euler a, CFG scale: 7, Seed: 1008289568, Size: 512x512, Model hash: 637d5dcb91, Model: DreamShaper332BakedVaeClipFix

SEAL



“Steps: 10, Sampler: Euler a, CFG scale: 7, Seed: 1654577914, Size: 512x512, Model hash: 637d5dcb91, DreamShaper332BakedVaeClipFix”

Noticing the curious results outputted from isotopies, that is, words with multiple or ambiguous meanings, I decided to write a new experiment. I wanted to be able to navigate around a concept and generate its different meanings. In order to do that, I took advantage of the mathematical properties of the models. Once a prompt is defined, it gets split into *tokens*, small units of symbols and letters that may or may not correspond to full words. ‘Powers that be’, for instance, becomes the sequence:

```
<|startoftext|>  
powers  
that  
be  
<|endoftext|>
```

These units have corresponding index numbers from a vocabulary, which in this case has 49,408 listings, becoming a sequence of numbers:

```
49806  
9422  
682  
655  
49407
```

Another operation then transforms this sequence into a collection of 768 floating point numbers. This is the latent vector, which encodes the ‘location’ of both the text and the corresponding image in the latent space. If this were a bidimensional map, it would be easy to imagine a circle around this coordinate: to explore this place, we could look into what is stored at points in a circle around it. But working on a space of 768 dimensions requires a mathematical trick, an algorithm proposed in 1972, that will help us obtain points on the surface of what is called an *n-sphere* (Marsaglia 1972).

In the grids below, the left image is the one generated by the prompt. The others are obtained by exploring an *n-sphere* around the coordinate. The code is very simple and was uploaded to my repository (Caldas Vianna 2023).

BOXER



seed: 10000002, num_steps: 25, sphere_radius: 5, model: Stable Diffusion 1.5,

TRUNK

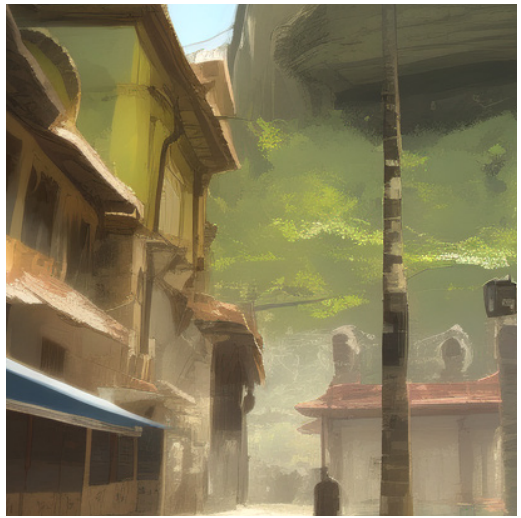
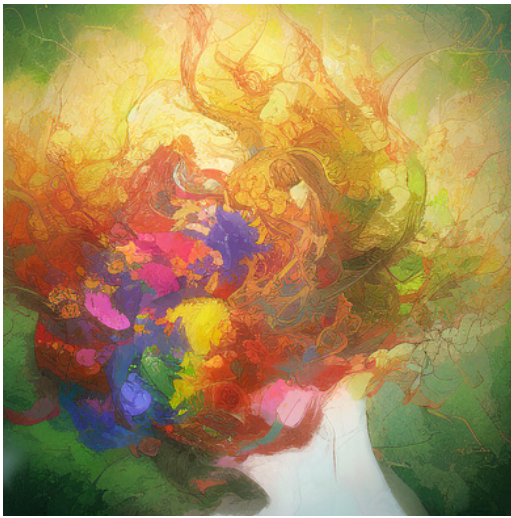


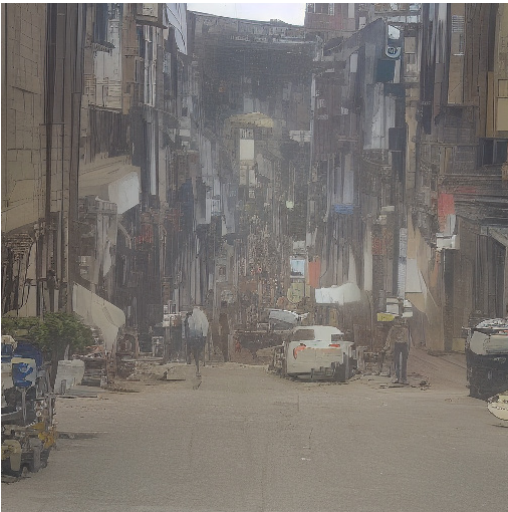
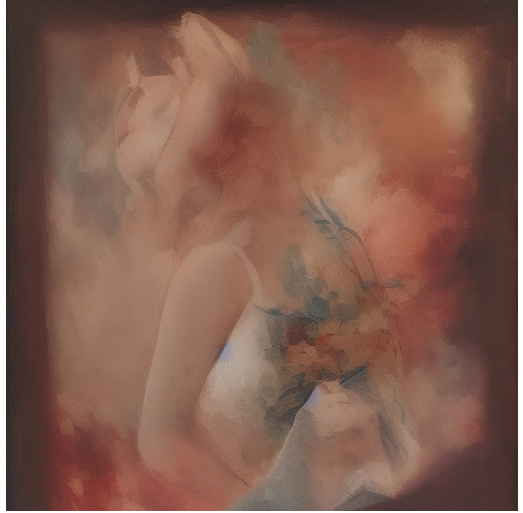
seed: 10000002, num_steps: 25, sphere_radius: 5, model: Stable Diffusion 1.4,

However, the latent vector is not the only variable that defines how the final image looks. As the parameters cited next to the images indicate, other choices include the number of steps, the sampler and the seed. This last one is important for us since it defines the random noise from which the diffusion process starts. Repeating the seed allows us to recreate the same image, from what is called a pseudo-random number generator. Asking for a true random number, or feeding it with a different seed, would generate a different image within the same semantic region of the latent space.

Voidism

A final experiment around text-to-image models is presented now. In *Voidism*, I propose images generated using Stable Diffusion where the text prompt is empty, that is, nothing is conveyed for the system to translate into images. For now, I suggest absorbing the qualities of the pieces while trying to fathom some sort of intrinsic transcendence.





Voidism is a cyborg arrangement. While it is true that the lack of text guidance allows for a much richer dialogue between the neural network and the human, beneath the apparent agency of the machine there are just too many options that had to be made by the human agent. For starters, there are quite a few choices about which of the existing trained networks to use, each with very different results. Stable Diffusion is the most used among the available networks that are free and public. I chose it more for its openness than aesthetic reasons, since it allows me to publish the exact parameters used and let them be replicated, a much-needed capacity in an academic publication. At the same time, it is known that some platforms like Midjourney modify the prompt, adding elements to make the images look better. In fact, Midjourney outputs have a style that is easily recognisable. So in order to have complete control over the prompt and be sure it would not be modified, I had to run my own model. But even within SD there are choices to be made: the first file released was 1.4, by Stability.AI, but there is also version 1.5 from RunwayML, besides versions 2.0, 2.1, and other files customised by users. I tried several of these before deciding on 1.4, which gave me the looks I could relate better to the concept of *Void*.

Within the same trained network file, other parameters must be chosen. The number of iterations in the diffusion process dictates how far the program must work towards obtaining a final result, usually between 10 and 50 loops. This parameter ended up being different for every chosen picture. Clip guidance scale is a number that indicates how closely the program must try to follow the textual description. Even in the absence of a description, the number has a great influence on the result. Many other choices are available, but probably the last (but not least) agential decision is the choice between millions of possible generated pictures. The initial noise, when the diffusion process evolves into a coherent image, can be left to randomness or picked from a seed value from 0 up to a few million. In my process I kept the seeds from the best random results to evolve them, changing other parameters, until I was satisfied.

In this description of the arrangement, I seem to be giving great importance to the by-products. In reality, endless different results could be found that would satisfy my criteria in making

a set of images related to the concept of voidism. The process also illustrates that the idea of a guideless algorithm is a fallacy. The empty prompt is just another specific location in the latent space, which I could use through heavy manipulation to develop a visual concept. Granted, the fact that the address is not related to a linguistic reference allowed for a greater freedom of outcomes. Compared to the development of the concept of *voidism* and my efforts to use it to communicate issues within this doctoral research, the resulting images might have less importance. I strongly believe that the arrangement is my creation, and it entitles me to some kind of authorship recognition. But I feel I would have a weaker claim towards any kind of authorship — let alone *ownership* — of the by-products.

In contrast, *Tactical Entanglements*, a book by Martin Zeilinger that resonates strongly with this research in many aspects, happens to make the case for authorless art. Zeilinger takes up the apparatus of copyright to discuss the problems of agency in times of artificial intelligence arrangements (Zeilinger 2021b). The cases of Michael Noll and Barrat/Obvious are also used by him to situate the problems brought by generative tools to the framework of copyright. But the paradigmatic case brought by Zeilinger is the work *All We'd Ever Need Is One Another*, by Canadian artist Adam Basanta.

First presented in a gallery in Montreal in 2018, the piece consists of two flatbed scanners mounted in a way that allows them to scan each other. The scanning surfaces, instead of digitising sheets of paper, are mounted in parallel and facing one another. Thus, when they are activated, they 'record distorted, unfocused light patterns that hit their glass surfaces' (Zeilinger 2021b, 97). This is all done without human intervention and with randomised parameters. The results are abstractions of light and shadows, generated at the rate of approximately one thousand per day.

The next step in the process is the selective filtering of images. In the words of the artist, if a scan is 'similar enough to a work that the art market or international collections have deemed art-worthy, then that image, which is similar to it, is also art-worthy. It becomes art' (Hannay 2018). This is achieved through a neural network specially trained with existing art pieces, which once fed with the scanned image will predict its similarity to artworks of relevance to the 'art

market or international collections'. Once it is found to be more than 83% similar to a work from the public set, the scan gets printed and receives a title that makes reference to the matched artwork and its author.

This arrangement obtained notoriety because one of its by-products was the subject of litigation. The piece named *85.81% Match: Amel Chamandy 'Your World Without Paper'* drew the attention of artist Chamandy, creator of the piece *Your World Without Paper*, who felt was having her copyright infringed by Basanta, and sued. The parts have settled, but the legal implication of the process was that the Basanta's authorship of both the arrangement and its by-products was recognised by the judicial system.

Zeilinger's thoughts coincide with mine in many respects. A summarising postulate that could perfectly encapsulate my own thesis can be found in it: 'the becoming-tactical of AI in critical artistic practice (is) a development that mobilizes AI's emergent capabilities for interrogating, exposing, problematizing, and challenging the aesthetic, ideological, or technological frameworks driving the commodification and propertization of creative expression' (Zeilinger 2021b, 27). I agree that *All We'd Ever Need Is One Another* is one of the finest examples of how AI allowed art to advance machine agency to territories never explored, and I too consider the flood of AI-generated content an unmissable opportunity for society to discuss the failed model of copyright used today.

However, the writer seems to find it necessary to model the creative assemblage allowed by AI as if human agency played no role: 'Functioning as an independent art factory, the project quite plainly cuts the traditional author figure out of the equation. [...] *All We'd Ever Need Is One Another* constitutes a posthumanist agential assemblage in which the human artist is participant rather than creator' (Zeilinger 2021b, 104). I agree that the ownership model of copyright is based on romantic fictions of authorship, but not because these new agential arrangements are proving that there is no human agency behind them. Firstly, these are artificial constructs because every creation is a derivative in the creative commons sense, meaning that no artist creates in isolation. It is impossible to avoid being influenced by previous works, not to mention non-copyrightable effects like news and opinions, and non-human

agencies. And secondly, without the artist's agency this arrangement would not have come into existence. By suggesting that Basanta is a participant rather than creator, Zeilinger, in my view, is writing another unnecessary fiction.

AESTHETIC ISSUES

As in with Botto and Abraham, voting-as-agency also appears unexpectedly in one of the most important pieces of the Stable Diffusion generative system — and possibly other generative AI models. As seen previously, Stable Diffusion was trained on a huge image-text pair dataset named Laion, composed of five billion images downloaded from the internet together with their textual captions. But the actual set used for training was not the original collection of five billion, but a smaller subset appropriately named Laion-Aesthetics. So how did aesthetics end up in big data?

Aesthetics judgements are notoriously difficult to automate. Computer art pioneer Friedrich Nake identifies two efforts in the early days of computational media to provide methods to evaluate aesthetic quality, one by Max Bense in Stuttgart and another by Abraham Moles in Strasbourg (Nake 2012). This aesthetic evaluation is not limited to art production, and in fact it doesn't place beauty, which is subjective, at the centre of the question, as Baumgarten defined it (Baumgarten 1750). Influenced by Shannon's Theory of Information, published shortly before their time, Bense and Moles set off to find mathematical functions that could beget an analytical understanding of the features of an aesthetic object. Since Moles was more focused on time-based media, music in particular, I will take a deeper look at the proposals of the former.

One of the inspiring grounds for this analysis was laid out by American mathematician G. D. Birkhof, who related concepts of order (O) and complexity (C) to define the aesthetic measure M (Bense 1998).

$$M = \frac{O}{C}$$

Thus, it reveals itself as a theory of organisation and simplicity, where greater orders of complexity would jeopardise the aesthetics

in question, and Birkhof explored this idea extensively on polygons. The refinement of this work was done in great part by these professors' PhD students. Rul Gunzenhäuser made the equivalency of complexity with Shannon's average statistical information H (Gunzenhäuser 1962). Helmar Frank brought the concept of surprise into these equations, again underlining the importance of creating outside the expectations of the system for a meaningful aesthetic contribution. By the end of the 1950s there was already a set of equations that could be used to evaluate a given image (Nake 2012, 8). This in turn was applied by students to perform a series of evaluations on what included industrial design objects and architectural façades in the late 1960s. It was also used by Nake himself during a residency at a computer centre in Toronto. There he reached the conclusion that in order to satisfy the equations, the output would be closer to 'a probability distribution: an infinity of images', which resonates strongly the infinite latent space of current generative AI methods. The same experience led him to dismiss the application of information theory towards an algorithmic aesthetics: 'This simplistic model was good enough for the study of traffic over telephone lines, but it had no relevance for the study of human communication. The essence of art is of a totally different kind' (Nake 2012).

Nake also notes that:

'If we adopt the position that aesthetics is concerned with sensual cognition, algorithmic aesthetics must by necessity reduce sensual cognition to those aspects, or approaches, that are standard, common, or average within a population and do not depend on individual living persons. A model agent must be assumed as the target' (Nake 2012).

Which leads us to the matter at hand: what was the aesthetic criteria — or algorithm — used by Laion to select about 15% of images from the original dataset that were used to train Stable Diffusion? As it happened, it was a voting system. Once more, generative systems turn to human agency to afford some kind of aesthetic resolve.

The documentation on the Laion website states that three different rating systems were used to filter images down from the big set. These systems did not select pictures directly, but instead the

voting results were used to train a neural network, named *predictor*, which was then used to select the most ‘aesthetic’ images from the five-billion large main set.

If the predictor is trained by humans, and since aesthetic judgment is a highly subjective trait, the question inevitably appears: who were the humans who voted? Whose taste was imprinted in the model that might be spawning images in the order of millions every day now? This required a bit of investigative work, which is not yet entirely concluded.

The creation of the Laion dataset is owed to a group of hobbyists led by Christoph Schuhmann, who in turn is not a data science professional, but a high school teacher working in Hamburg (Rainer et al. 2023). Some information gathered here is the result of my direct interaction with Schuhmann in the Laion Discord chat environment.

While they were gathering (*scraping*, to use big data jargon) their large collection, another digital tinkerer named John David Pressman⁵⁸ was concerned with how to create ‘aesthetic models to guide AI-generated imagery into better satisfying human preferences’ (Pressman 2023b). For this purpose, he programmed a bot that ran on the Discord platform. This contraption would display AI-generated images to the users, who would rate them using a scale from 1 to 10. The results were anonymised, stored and made publicly available on a database (Pressman 2023a). Unfortunately, Pressman did not reply to my enquiries, and I couldn’t identify in which context this data was collected. Discord bots must always run on Discord servers, which are usually associated with some institution or group. It is the most popular interaction platform among the visual AI crowd; therefore open servers exist for Stability.AI, Disco Diffusion, Midjourney (the user interface of which is built on such servers), OpenAI and others. Apparently, there used to be a *Simulacra* Discord server and its participants could cast votes, but I could neither confirm this fact nor gather who these voters were.

Nevertheless, Pressman’s motivations and the documentation he left make it clear that he is also concerned with the bias that emerges from this type of data filtering. *Simulacra*’s code repository

58 <https://web.archive.org/web/20221207015914/http://www.jdpressman.com/>, accessed 29 August 2023

identifies some of these issues, like the fact that it did not show copyrighted content (therefore limiting the visual experience of voters), did not show material deemed not safe for work environments (eschewing nudity), and there were no hate speech-associated images. In fact, an onboarding test was performed on participants to estimate whether their own bias would interfere with their aesthetic judgment. This was achieved by displaying a set of pictures that contained, for instance, a sharp, detailed painting of Adolf Hitler. A low rating on what was considered, by the developer of the ratings platform, as a well-executed depiction of a despicable character could indicate a tendency to vote for political rather than aesthetic criteria. On a personal note, I am not sure whether such judgments can be evaluated separately — and if anyone can be the judge of aesthetic values. But that was Pressman’s method.

Even not knowing where the voting was done, the documentation makes a clear statement on the characteristics of its constituency:

‘Participants Are WEIRD

Participants in Simulacra Aesthetic Captions are largely WEIRD, or Western, Educated, Industrialized, Rich, and Democratic (Henrich, Heine, and Norenzayan 2010). This means that the aesthetic preferences recorded are not universal among humanity. While we didn’t take a survey of the demographic makeup of SimulacraBot users, it should be assumed they largely reside in the United States and Europe. We welcome replications of Simulacra Aesthetic Captions which sample their results from other locales and demographics.

Users Are Mostly Open Source AI Developers And Enthusiasts

Further narrowing the scope of aesthetic feedback in Simulacra Aesthetic Captions is the userbase consisting largely of people who are power users and developers of AI art. This means that their aesthetic feedback is going to lean STEM, fantasy, nerdy, esoteric, etc’ (Pressman 2023b).

And lastly, but not least relevant, Pressman notes in a tweet that AI-generated imagery has deficiencies that can turn to new bias (Pressman, John David [@jd_pressman] 2022). A known shortcoming of current AI models is their failure to generate hands and feet with the correct number of fingers and toes. A low rating on such images — which in this case represent most images with hands and feet — would be translated into a prejudice against hands and feet in general, jeopardising the appearance of these limbs in the models trained by this predictor.

Regardless of the deficiencies within the Simulacra database, the set was picked by Laion to train their predictor. In other words, the 176,939 ratings of images were used to train an artificial curator. But other datasets were used too. Schuhmann noted that the images in Simulacra did not have examples of pictures with texts on them, which could lead to a problematic bias against logos. To address this, he created a set with 15,000 such images, the Laion-Logos dataset, and asked his high school students to rate it — again, sampling the aesthetic values from a very limited slice of the population.

Finally, the Aesthetic Visual Analysis dataset is also used by Laion. This set comes from an online Digital Photography competition where users can cast votes on their favourite photos — <https://www.dpchallenge.com/>. Although tending to a broader audience than visual AI geeks and German teenagers, this population sample is also limited to photography hobbyists with internet access, and it is likely to have many intersections with the above-mentioned group of WEIRD people.

Armed with these criteria, Laion's programmers were able to create a mechanical curator that would have made Max Bense and Abrahams Moles envious. This predictor created its own automated ranking of the five billion images in the original Laion-5B set, of which 600 millions were graded higher than 5. This subset above the grade point cut-off was used to train Stable Diffusion models versions 1 and 2, which, since being released as open access, became the most popular generative visual artificial intelligence tool by the end of 2022.

The repository of images in the dataset and their rating can be seen here: http://captions.christoph-schuhmann.de/aesthetic_viz_laion_sac+logos+ava1-l14-linearMSE-en-2.37B.html. Some of the

images might not be available anymore, since the URLs in the dataset might point to defunct websites. An archived version exists in the Internet Archive: https://web.archive.org/web/20220906173111/http://captions.christoph-schuhmann.de/aesthetic_viz_laion_sac+logos+aval-114-linearMSE-en-2.37B.html

Needless to say, from a global perspective, this is a very limiting, even problematic, set of criteria. In a way, it serves only its own kind — the same people who trained the models with their subjective taste are the ones who use it the most. But this aesthetic judgment excludes a large proportion of the visual culture of the world, in particular people who are not interested in generative AI or digital photography. The graphic taste of, say, Indonesian housewives, African street artists and Caribbean naif painters are lost in such models, condemning the software to a limited field of view like a horse's eye mask.

This is not just about the loss of the models, but also the largest part of the population that cannot see itself represented in the repertoire of tools that might come to heavily influence visual culture in future years. It is a type of neural colonialism, which works twofold. On one hand, by barring non-Western visual traditions from future artefacts of cybernetic imaging, eliminating vast extensions of graphical culture. And on the other hand, by the visual extractivism it performs on the production of human artists — from the Global South or not — by using their materials to train generative visual neural networks. This resonates with a broader criticism of artificial intelligence on the exploitation of labour used to train it — like the Kenyan workers making less than two dollars an hour to exclude disturbing content from the training dataset of OpenAI's ChatGPT (Perrigo 2023). Ethical AI researchers have already warned of the perils of large language models that *seem* coherent and truth-ground; the same concerns can be applied to generative images that seem to respect aesthetic yearnings but are created from a very particular judgement (Bender et al. 2021).

The Laion-Aesthetics set serves as evidence that no matter how capable today's computers are at creating compelling images, they are still unable to make a judgment without human guidance. Like with the Botto and Abraham projects, agency is delegated to humans, in the form of data obtained through the voting process.

As seen, the suffrage in these processes is less than universal and leads to specific bias problems pertinent to each case.

Our imaginary autonomous artist, Darci, would not have such shortcomings. It would develop a coherent aesthetic of its own, in resonance with human visual culture, making relevant contributions to the art world.

At this point I should disclaim that after contacting Laion regarding these problems, I was invited to help mitigate them. As Laion is a not-for-profit, open institution, I promptly agreed to volunteer, and I hope I can make contributions with positive effects on the next models.

Art against AI

The appearance of artificial intelligence systems that can generate high-quality images from textual descriptions created a strong backlash from a significant share of the creative visual sector.

The arguments against AI art usually come in two shapes. The first is a re-enactment of questions that always appear when a new technology becomes available. One of the best-known historical examples are the Luddites, a faction of organised textile workers in 18th-century England who destroyed industrial equipment. These actions were a form of protest against the loss of their jobs, which were being replaced by mechanical looms. It is not difficult to imagine that it was not the first time similar concerns appeared. The advent of the printing press certainly had an impact on the activities of manual book copiers. Closer to this research, Norbert Wiener's founding book on cybernetics conveyed a discussion on the possible loss of work and conversations with union leaders (Wiener 1948, 28).

Like in most previous cases, the risk of manual artistic jobs being replaced or at least reduced by AI is an actual threat. The same concerns are happening among software developers and in the editorial industry, as systems like Co-pilot and GPT-4 are able to produce commercially acceptable code and writings. But loss of jobs is not the sole effect of new tools. The creation process will change, as artists incorporate these mechanisms into their process. And newer functions will appear, requiring a specialised workforce to deal with these new potentials.

I suggest calling this new breed of anti-technologists as *neuraluddites*, for their main adversaries are neural networks. Now, it should be clear that I consider the Luddites a vector that is both reactionary and revolutionary. Their struggle against technology may seem quixotesque, as it is practically impossible to constrain mass adoption of an advancement with so many benefits in the capitalist sense, and also in the communal sense. But they also force society to rethink the consequences of this adoption and find solutions for them.

The second argument against AI art is not based only on labour dangers, as the pace of technological evolution has shown to be historically impossible to slow down. Instead, it stems from the fact that every neural network requires massive amounts of data to train from, and this data comes mainly from humans who are never credited, let alone rewarded, for this input. Once again, ownership and authorship claims have evolved to the point of litigation.

One of the loudest voices against the use of AI in art belongs to illustrator Karla Ortiz⁵⁹. After experimenting with generative web platforms like Lexica⁶⁰ in September 2022, she reacted very strongly against⁶¹ it in the commentary of an interview with Stability.AI CEO Enid Mostaque. At the beginning of October, she was sharing anti-AI opinions several times a day on the Twitter social network.

‘[...] I know AI companies like Stability AI, knowingly and purposefully took and used visual artists and general public copyrighted data and private data without a care or concern (aside from seeing data on where they scraped their data from.)?’ (Ortiz [@kortizart] 2022)

In a blog post in December, Ortiz made various claims which, although not wrong, do not support her own argument against the use of AI tools for generative purposes: ‘Looking at imagery of others is a part of learning how to draw/paint, but unlike AL/ML models, looking at images of others isn’t THE singular way we artists learn

59 <http://www.karlaortizart.com/>, accessed 14 June 2023

60 <https://lexica.art>, accessed 14 June 2023

61 <https://twitter.com/kortizart/status/1569836239247003651>, accessed 14 June 2023

or create' (Ortiz 2022). Building on her point, I would say that the images generated by AL/ML models do not result exclusively from the computer, but from a collaborative arrangement where agency is distributed between the artists and the tools. They are not only guided by the prompts, but also from varied parameters, and more importantly, from a long iterative process of trial, error and selection of the results. Only after that, these pieces arrive at a point that can be shared or used.

In January 2023, Karla joined artists Sarah Andersen⁶² and Kelly McKernan⁶³ in a class-action suit⁶⁴ filed by litigators from a law firm already known for suing⁶⁵ Microsoft on account of the AI programming assistant Copilot. Their main argument pleads that text-to-image AI models are the '21st-century's collage tools'. There is scepticism from other copyright litigators about the allegations. I would like to incorporate here a quote by the former president of the Los Angeles Copyright Society, where he expresses several points of relevance from a point of view that is very close to mine.

'While it is true that using copyrighted images to train the models is not a granted right within the legal framework of content ownership, other businesses like Google Books do worse by actually displaying a portion of the copyrighted material to the clients.

Copyright owners have the exclusive right to make derivatives of their work. However, while the definition of derivative is far from closed, they should "incorporate a portion of the underlying work", which the images generated by AI processes in question do not. " (Moss 2023).

Finally, regarding the use of other artists' style, the lawyer suggests that while "Style is certainly an element that can and should be considered within an overall substantial similarity analysis, (...) prohibiting works that are merely 'inspired by'—or even copy—preexisting

62 <https://sarahcandersen.com/>, accessed 14 June 2023

63 <https://www.kellymckernan.com/>, accessed 14 June 2023

64 <https://stablediffusionlitigation.com/>, accessed 14 June 2023

65 <https://githubcopilotlitigation.com/>, accessed 14 June 2023

art techniques would artificially stifle human creative development. [...] The courts that have considered this issue have held that style is an ingredient of expression, but that standing alone, it isn't protectable" (Moss 2023).

Important institutional voices in the open culture domain have manifested their support for the use of copyrighted images for training AI. Creative Commons, the copyright think-tank that enabled the widely used CC licences, generally affirms that the right to use mined data depends on the use, but also states that 'At CC, we believe that, as a matter of copyright law, the use of works to train AI should be considered non-infringing by default, assuming that access to the copyright works was lawful at the point of input. For example, (text-and-data mining) in the context of research or education should be allowed under an exception to copyright, following the adage that "the right to read is the right to mine"' (Vézina and Pearson 2021). The Electronic Frontier Foundation, a leading entity in defending civil liberties in the digital world, believes that 'Done right, copyright law is supposed to encourage new creativity. Stretching it to outlaw tools like AI image generators—or to effectively put them in the exclusive hands of powerful economic actors, who already use that economic muscle to squeeze creators—would have the opposite effect' (Walsh 2023).

In the meantime, the actors accused of acting maliciously in using copyrighted data are moving to create tools and models that allow content owners to opt out. A start-up named Spawning AI has created a tool that allows illustrators that post images in sharing platforms such as Deviantart and Artstation to opt out of being used for training. Stability AI, which trains the Stable Diffusion models, said it will abide by these options. Until the deadline of 3 March 2023, 80 million images have been tagged for removal from the sets (Kemper 2023).

Unfortunately, the opposition to AI generative imagery has reaped the support of scholars with very relevant works around the problematic issues of AI ethics in its applications in society. Timnit Gebru refers to Ortiz's blog post to state that AI-based visual work is not 'inspired' (Gebru [@timnitGebru] 2023a). She also praises the development of tools like *Glaze*, which supposedly empowers the enforcement of the rights of bricks-and-mortar artists (Gebru [@timnitGebru] 2023b).

Glaze is an image filter published by a group of researchers, which works by adding data that would spoof the training process, generating errors. The technique became known as *poisoning* the dataset. The software was immediately dismissed by programmers as too specific and easy to spoof (Illyasviel 2023; Shan, Cryan, et al. 2023). A new version, named *Nightshade*, was released in 2024, and it is supposed to be more effective (Shan, Ding et al. 2023). A few issues might still hinder its deployment, though. The process of preparing the images is very intensive in terms of computation. In order to have an effect on the foundational dataset, it would have to include many poisoned images, in the order of tens of thousands. Since, as explained, the process of training a new model is very expensive, it is unlikely that the concept will be tested within an actual training soon. Models like Dall-e 3 have been using synthetic images in their training (Betker et al. 2024). The code for the tool is not open, only the tool itself, which is another reason its effectiveness cannot be independently confirmed. Besides, poisoned images can be tagged and avoided in training sets. The tools do not solve the problem of already trained models, which have been downloaded by endless users on the internet. Since there are several datasets that have been built from images scraped before the existence of these tools, there is nothing to stop them from being used in new trainings. In any case, it is still too early to assess the effects of these tools in the generative visual AI landscape.

So far, the paradigmatic case for copyrights in AI-generated work is the comic book *Zarya of the Dawn*, by Kristine Kashtanova, who refers to herself as a 'prompt engineer'. Kashtanova filed the work for copyright protection in the United States in 2022, without disclaiming its images were created using Midjourney, a text-to-image platform, and was granted registration. However, when the United States Copyright Office learnt that the comic used machine-generated images, it announced the registration was being reviewed. The final decision came out in February 2023: the graphic composition of the comic book, as well as the text and story, were under protection. However, the images were not (Lindberg 2023). This was due to the fact the author must be a human: the 'U.S. Copyright Office will register an original work of authorship, provided that the work was created by a human being' (USCO 2021). Nevertheless, other legal

jurisdictions such as the United Kingdom have a different understanding: ‘In the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken’ (UK Copyright, Designs and Patent Act, 1988, section 9(3)).

The legal landscape in the European Union is not so clear, as at the moment of writing there is neither a common directive nor enough jurisprudence. Senja Assinen speculates that AI-generated materials would not be entitled to copyright, given the requirement of a human creator (Assinen 2021). However, it is difficult to be precise about how much human input — like a prompt — would be the threshold to allow for protection of such works. Surveys of the Finnish (Kekola 2020) and Spanish (Calleja Reina 2023) legal frameworks arrive at similar conclusions.

Even though I have been using copyright extensively as a standard against which to measure authorship, its reach within society is not absolute. While it is important to major cultural industries such as Hollywood movie studios, the music industry and publishers, the art world has its own authorship conferral mechanisms. For the auction house Christie’s, the fact that *Edouard de Belamy* could not be copyrighted was not an issue; nor was it for the collector who spent more than USD 400,000 on the piece. Even performances can be negotiated, using other types of legal assurances, like contracts and notarial certificates of authentication. The same logic applies to NFTs, which actually serve as databases of ownership, rather than any type of anti-piracy mechanism.

At the other end of the spectrum of cultural industries, independent artists working in small communities or in the cultural context of the Global South also have no means to enjoy the protection of copyright control systems. They may survive by immediate mechanisms such as grants, direct sales and live performances, which might never achieve enough massification to reap the benefits of economies of scale, but might just as well generate enough revenue for their subsistence.

As for this research, the struggle for copyrights in generative works helps me frame the post-human dispute in agential arrangements. While I strongly think that such arrangements are human

in their origination, the stance of institutions such as the copyright office, and of thinkers and artists like Gebru and Ortiz, show that the belief in some kind of non-human authorship is growing strong.

The *autopoetic aesthetic arrangement* framework proposes that fixed-form, copyrightable products should be given lesser importance than the system built for their generation. As such, I think there is a good argument *against* their copyright, even because in some arrangements it would be physically impossible to register every possible result. The grounds for generative art copyrights should not be that there is no author, but that what is worthy of authorship are the methods and data used to generate content. Allowing by-products to be free of rights will lead to an explosion of a free, remixable culture that can reshape creativity.

The case for copyright-free generative work is even stronger when it uses AI models trained on a considerable slice of humankind's historical visual culture. These data-intensive networks are built on a shared heritage that should be treated as part of the cultural commons, the universal cognitive knowledge, which is more than available for all, in fact it is unavoidable. These collective resources should not end up in the hands of the few with access to copyright enforcement mechanisms and rights to vast cognitive material. The legal framework of some countries, like Japan, understand that the rights of the author do not limit the use of their production to inspire new works, remix the originals and, more recently, train neural networks with them. The EU data protection regulation opens up an exception to the use of private data for scientific purposes, which allowed the first version of the Stable Diffusion model to be trained in Germany. This stance, I believe, should be adopted globally; in fact, having just one country allow this is enough to hinder the enforcement of such a rule, since models are trained and hosted in computers accessible worldwide. Nevertheless, there could be exceptions to this concept of global commons, for instance to make exemptions to the use of indigenous cultures, if we considered these to exist outside this globalised network of creation.

I should also clarify that my position is limited to the use of materials in training in ways that creations cannot be reused directly; in other words, the original artwork will not appear as the result of algorithmic generation. Although I am a firm believer in the

adoption of more permissive licences such as Creative Commons, the decision to authorise the reuse of fixed works should be individual. Reuse can be concerning in cases where one's material is given political uses that go against the creator's beliefs, for instance. My concern is that, as technology changes and content corporations become more powerful, the reappropriation of *styles* is becoming more limited. In the context of the United States judicial system, this is already becoming a reality: in 2015 musician Marvin Gaye's estate won a lawsuit against Pharrel Williams and Robin Thicke, proving that the late soul singer's songs 'feel' were copied — even if no part of the original melody, harmony or even basslines were used (Zernay 2017). The decision caused a flood of similar lawsuits, as until then the remixing of styles was not considered an infringement of copyright (Elbeshbishi 2022).

My suspicion, though, is that big copyright owners will soon realise that there is money to be made in these assets. They might use lobbying influence on law-making to continue exercising their dominance on contents and markets, pushing towards legal changes to expand the limitation of creative appropriation.

There is a third factor in the reaction towards generative AI tools, namely the potential to produce convincing forgeries. This is probably the most profound and consequential issue for society at large. Even before the advent of visual AI, election outcomes have been manipulated by fake news and propaganda. For malicious actors, the possibility of creating fake but photorealistic images with public figures represents a powerful new toolbox. Countries with more control over the expression rights of their citizens, like China, are dealing with this by requiring an 'obvious mark' on every AI generated image (Edwards 2022). While it is not possible to cover the issue here as the implications are too vast for a thesis focused more on the possibility of autonomous creative agency, my concern about it is hereby dutifully noted.

Societal issues around artificial intelligence

Authorship, labour and creativity are far from being the only issues brought about by the recent rise in artificial intelligence. Many of the big neural systems are black boxes causing environmental concerns

due to the intensive use of electricity and water in data centres, both at the time when models are built and when they are used. They also bring to bear problematic relations in their training phase, since low-wage workers from developing countries are hired to evaluate content that is often violent and disturbing. Besides these visible issues, opening their black box can be a challenge in itself, since some of the most influential companies do their development without sharing their methods, code or models.

As its name suggests, OpenAI was founded in the spirit of altruistic research, with the goal (previously discussed in this thesis) of mitigating the risks to society that artificial general intelligence could represent (Introducing OpenAI 2015). And they were committed to sharing the results of their findings:

‘(our) Researchers will be strongly encouraged to publish their work, whether as papers, blog posts, or code, and our patents (if any) will be shared with the world. We’ll freely collaborate with others across many institutions and expect to work with companies to research and deploy new technologies’ (Introducing OpenAI 2015).

At this moment, both the non-profit and open stances seem to have been forgotten. For instance, their most known invention is the GPT series of large language models. But they only published their weights (the trained files used for generative tasks) until version 2 (GPT-2 2019). GPT-3 had only a description published, while the ground-breaking GPT-4 still has its details kept in secrecy (Brown et al. 2020; Heaven 2023). As mentioned previously when discussing John Searle’s *Chinese Room* concept, this makes scientific research on these models very limited. It also hinders efforts to calculate the environmental impact of training the models. Finally, commercial secrecy makes the evaluation of the use of ChatGPT (which is based on GPT-4) impossible. Therefore, only imprecise estimations of these can exist. An estimation of the training emissions for Bloom, a model with 176 billion parameters, pointed to up to 24.7 metric tonnes of carbon being used, equivalent to the energy consumption of 18 houses in one year (Luccioni, Viguier and Ligozat 2022). But GPT-4 could easily be one hundred times hungrier.

The human labour costs are equally difficult to measure. To curb unethical and even racist replies from ChatGPT, a technique named reinforcement learning from human feedback (RLHF) is used (Ramponi 2023). This requires that the outputs of the machine be read and classified by humans. A Time magazine investigation revealed that OpenAI hired Kenyan workers for less than two dollars an hour to do this toxic job. ‘Some of (the texts) described situations in graphic detail like child sexual abuse, bestiality, murder, suicide, torture, self harm, and incest’ (Perrigo 2023).

Image-generative AI might even use large language models as part of the creative process (to create text prompts, for instance). But the models cost much less to train and are mostly published in the open. The ground-breaking Stable Diffusion model cost the equivalent of 12 metric tonnes of carbon dioxide in emissions to train, similar to the consumption of 2.2 houses in one year (Rombach and Esser 2022). The generation of pictures is done in a matter of seconds on users’ domestic GPU-equipped computers, making a much less intense use of electricity than running video games, for instance, which may use the same devices for many hours.

CONCLUSION

Wrappings and unfoldings: the tortuous path to the poetics of autopoiesis

This project started with the naïve assumption that developing an autonomous artist, such as Darci, was a task not very far in the evolution of generative systems. This illusion was quickly dissolved in the face of my investigation around what constitutes a symbolic machine, and what are the human traits that give us a creative impetus. While it is surprising to see many prominent researchers, politicians and businesspeople share their concerns regarding an upcoming artificial general intelligence, I stand assured that these anxieties are more based on wishful thinking, fearmongering and morbid desire than they are grounded in reality.

The personal delusion described above is not a reason for any kind of frustration. It is instead a rewarding outcome, considering every research question must convey the possibility of being proved wrong. What is more surprising is the realisation that the *poetics of autopoiesis* were hiding in plain sight. On one hand, the textual tools available for poetry became available, even if in a tortuous way, to visual practitioners, through the emergence of semantic imaging tools. But perhaps more generally the research made me realise that within every creative system with autopoietic traits lies a poetic potential that resides exactly in its autonomous qualities.

Understanding that the systems that so often appear in the practices of art and technology projects all entertain some kind of autonomy made me realise that their poetic quality is deeply connected to the exact quality of being autopoietic. They create their own self-contained productive arrangement limited by boundaries in their relation to the artist, audience, and the elements it is exposed to, which can be incorporated or not in each specific design. The realisation would not have happened if it weren't for this deep dive into symbolic computation, systems theory and neural networks. With new types of autonomies, new poetics emerge.

My research in creative systems is informed by my own artistic projects, making it easy for me to identify how their expressive and aesthetic qualities relate to their autonomous traits. Seeing the current semantic imaging systems through the optics of autopoietic aesthetic arrangements is fundamental to understanding their complex relations. It is easier to grasp the role of endlessly generated pictures if we see them as *by-products*, or *allopoietic components*, rather than the final expressive work. In this configuration, the whole text-to-image generative system, in which the prompts assume a predominant role, takes precedence over the outputs. Prompts alone are also not enough to absorb the auratic nature of the arrangement — they are a piece among settings, parameters, models, aesthetic ratings, and the programming code needed to deploy this gigantic collaborative cognitive device. They also rely on the millions of imagetic and textual references used to train the models.

The evidence that this apparatus leans heavily towards a conceptual and procedural manifestation (instead of material), is that the entire arrangement can be forked. Armed with the data described above, one can take it as its own and adapt it from this starting point, to develop their own creation. Since the originality of the artwork resides in the combination of system and prompt, instead of fine motor skills, it is no wonder that some of the most recognised semantic visual artists keep their prompts and settings secret — much like industrial secrets of unpatented inventions. This is the case, for instance, with Arno Coenen, represented by the Dead End Gallery in Amsterdam, announced as the first AI gallery in the world. Having access to the prompt and tools would imply the easy replication of the works or at least the styles developed, as no dexterity is required.

The *arrangement* insight came about in the last stages of the research, and it unfolds a fan of rich paths for theoretical explorations. The concept needs to be held up against several thinkers who probed the connections between technology and creativity and were not part of the original research. These include Vilhém Flusser, Friedrich Kittler, Bernard Stiegler, Gilbert Simondon, Paul Virilio and even Walter Benjamin, just to name a few. The original research proposal that expressed a desire to relate thermodynamics, complexity and emergence remains unexplored; there's great potential

to rethink Prigogine and Stengers' contributions, particularly in face of the diffusion techniques within semantic imaging systems.

Important writings on the particularities of written language could not be addressed, but will be absolutely made more interesting in face of the developments in semantic generative AI. Especially relevant approaches include Lyotard's take on language games, Baudrillard's concept of *Difference*, and Foucault's epistemic study of representation that can be found in *Les Mots et Les Choses*.

Several of the discoveries in the doctoral project fit under a conceptual umbrella that can justify the organisation of an entity to embrace investigation of the problems surrounding semantic visual aesthetics, and, in many cases, their mitigation. As the doctoral project winds down, I am in the process of envisioning ways to continue these enquiries. Among my tasks, I would like to:

- Continue to investigate, through computation and statistical methods, the biased deficiencies caused by the excluding scoring and prediction systems used in training AI visual models, described in the previous chapter. These probings must refer to existing theoretical reflections on taste, such as Bourdieu's *Distinction* (Bourdieu 2002).
- Understand the role of visual AI in the discussion, started by Harun Farocki, around *operational images*. Although absolutely tempting, the idea of images, the relevance of which lies in their use within mechanical, societal and political operations instead of human appreciation, is a path that I could not explore within the ambitions of this project. Yet, visual AI offers a new paradigm for *operation*: the machine learning models swallow pictures and extract their conceptual essence, making them invisible in the process. As noted by Jussi Parikka, these technologies (and others) 'should be placed in historical dialogue with questions of data, sensing and spatial use of images' (Parikka 2023).
- Expand existing research on gaps and biases in gender, ethnicity, north/south relations, encompassing theoretical methods through decolonising thinkers like Achille Mbembe, and the fundamental deconstruction of discourses proposed by Jacques Derrida, together with analytical data investigation.
- Propose, develop and publish open access models and embeddings to mitigate such effects.
- Continue the mathematical exploration of meaning in visual latent

space, incorporating more semantic tools such as antonyms, synonyms, ellipses and recursion, while resourcing classic semiotics frameworks including, but not limited to, Umberto Eco's.

Beyond Darci

By following the arguments of anti-AI artists to the letter, it would seem as if the Darci-like artificial artist already exists. Their claim is that machines elaborate the images, without human interference. If, as claimed, all the work of creating an image in the text-to-image spawning processes is made by the machine, we should be worried that AI systems will take over human protagonism in the creative domain.

This claim is only partially true. In the arrangements, there is a displacement of agency towards the machine. This is evident in the struggles that happen in the process of creating images in these platforms. The language affordances of models imply that in order to obtain a desired effect, creators must fight uncontrolled outcomes that result from the inaccuracies of language. Words replace the brush as the extension of thought, as the tool used to convey imagined surfaces into fixed ones. Hence the long prompts, which try to be more and more specific and avoid misunderstandings in this cyborg communication. This was my point when showing the language affordances of semantic imaging models. What might be seen as an infuriating limitation of this arrangement — when one is fruitlessly trying to obtain specific results — should be seen also as an opportunity for poetic exploration. The rich and complex tools available for poets — polysemy, metaphors, analogies — had never been so close to the reach of visual practitioners.

The case of *Voidism* seems to erode the efforts towards human control of the semantic imaging process and reinforce the post-human argument. If no words are given to the system, and it still spawns out beautiful pictures, then the human input is irrelevant. From a transcendental perspective, it would seem that the ghost in the machine is creating images that convey its own mechanical subjectivity. It *decides* whether to make landscapes, interiors, cityscapes, portraits or abstractions. And worse, it seems to have developed a specific, 'personal' style. There is a coherence in the set exhibited.

It seems that a Darci-like arrangement might not be so distant if started from these affordances. Just like my rendition of the autonomous artist, it was also trained on millions of cultural visual references. But as we see, these artworks came into existence thanks to long interactions between author, results and parameters. To borrow a powerful image from Lambros Malafouris (Malafouris 2008), my agency moulded the images like clay, but the model sculpted back my will by proposing surprising or disappointing outcomes, or a cyborg collaboration on Donna Haraway's terms (Haraway 1991).

Looking back at my artistic trajectory, I can see how much I also looked for a cyborg poetics of autonomy, without success. My 2008 feature film, *Ressaca*, started as an investigation into automated movie editing with its roots in a project done ten years earlier, during my master's degree studies. Nonetheless, it ended as a performance where the agency was shared with the audience while I exposed the process of putting the story together in front of it. In 2018, I started an as-yet unfinished feature project, *Dispersão*, where the viewers would guide the narrative by interacting with a social network algorithm. The project faced the same issues: all experiments in algorithmic montage turned out to be inferior in comparison to the intuition and know-how of the human editor. Against these experiences stand the projects made during the doctoral research.

Photosynthesis is the most autonomous artificial image creator that I can conceive. Despite the original human impulse, consisting of my design and programming, it will continue to draw for the duration of the time its parts hold together. It is an autopoietic poetic system with a closure defined by solar energy intake and the output of tree pictures that may reverberate with viewers, if they take part in pointing the work towards the sun, or with nature and itself, in case it rests on a naturally lit spot. But I was responsible for conceiving it. The images it spawns are mere by-products of the arrangement that infers to the machine a specific agency in using randomness to create a combination of pixels that humans can recognise. As its author, I cannot consider it a spontaneous emergence of creativity.

Let us hypothesise a fiction where this scenario would come true: a creator that does not derive from human agency. Our Darci would not suffice, as its starting point is a human project to create an artistic robot-like arrangement. Calling it Super-Darci is not appropriate

— this prefix is too close to derailed transhuman dreams of an evolved race, and the fiction too far from our almost within-reach artificial painter. Let us just give it another neutral moniker like Mone.

Mone differs from the current autopoietic arrangements in that it does not originate from humans, and hence it is devoid of human agency and anthropomorphic cognition. As such, it could not be built; it would have to emerge. It is best not to speculate on the technology that would allow this to happen. I don't think computers as we know them today, or any Turing machine, would be up to the task; it seems to me that instruction-based and discrete devices would not be appropriate for a device that operates on the uncertainty of the physical domain.

Mone is an autopoietic aesthetic arrangement taken to its limit. It possesses no trace of human agency, and it is not allopoietic in that it is not contingent on an observer, and it won't output components that aren't meant for its own maintenance. It will be creative, but by fathoming a non-human subjectivity. If it can establish a dialogue with human culture, I will engage with it like an original artist; otherwise it will exist on an isolated domain, having little relation to human culture. If it is based on biological principles, it might be seen as a new species; otherwise it will be based on supports yet to be created or repurposed.

Computers are mini-mones, in that they have their own cognition. But since we humans designed and programmed them, we develop them as such that their by-products are meaningful to us, and such that we share a common symbolic cognition that allows us to create programs for them. The object of computer operations is data, the known. But 'In order for any intelligence to produce art, its object cannot be the known, but the unknown' (Hui 2021, 243). Asking for creativity from a state machine is like asking a Rubik's Cube for innovation. It just belongs to a different domain.

Mone will have its own cognition. As such, it might be useless for humans, in that we won't understand it. It might be an incomparable artist in its particular domain, but in expressions unreachable to us, like the greatest poet of a language we do not speak. At the end of this project, this is the best aphorism I can offer: If we one day succeed in building a completely autonomous, autopoietic, creative arrangement, it will probably be inaccessible to us.

References

- Abu-Khalaf, Murad, Frank L. Lewis, and Jie Huang. 2006. "Policy Iterations on the Hamilton–Jacobi–Isaacs Equation for State Feedback Control With Input Saturation." *IEEE Transactions on Automatic Control* 51(12): 1989–95.
- Adamatzky, Andrew, and Theresa Schubert. 2014. "Slime Mold Microfluidic Logical Gates." *Materials Today* 17(2): 86–91.
- Adams, Kevin, Patrick Hester, and Joseph Bradley. 2013. "A Historical Perspective of Systems Theory." *III Annual Conference and Expo 2013*: 4102–9.
- Adorno, Theodor W., Gretel Adorno, and Rolf Tiedeman. 1997. *Aesthetic Theory*. Minneapolis, Minn: University of Minnesota Press.
- Agüera y Arcas, Blaise. 2022. "Artificial Neural Networks Are Making Strides towards Consciousness, According to Blaise Agüera y Arcas | The Economist." <https://web.archive.org/web/20220609093229/https://www.economist.com/by-invitation/2022/06/09/artificial-neural-networks-are-making-strides-towards-consciousness-according-to-blaise-aguera-y-arcas> (25 January 2023).
- Alom, Md Zahangir et al. 2018. "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches." *arXiv:1803.01164 [cs]*. <http://arxiv.org/abs/1803.01164> (21 September 2021).
- Arendt, Hannah, and Ronald Beiner. 1992. *Lectures on Kant's Political Philosophy*. ed. Ronald Beiner. Chicago: Univ. of Chicago Press.
- "ASC: Foundations: History of Cybernetics." <https://www.asc-cybernetics.org/foundations/history2.htm> (18 October 2022).
- Ascott, Roy. 1968. "The Cybernetic Stance: My Process and Purpose." *Leonardo* 1(2): 105–12.
- Ashby, W. R. 1947. "Principles of the Self-Organizing Dynamic System." *The Journal of General Psychology* 37(2): 125–28.
- Asimov, Isaac. 1950. *I, Robot*. Garden City, NY, USA: Doubleday & Company, Inc.
- Assinen, Senja. 2021. "European Union Copyright Protection for AI-Generated Works." Master's thesis, University of Turku. <https://pdfs.semanticscholar.org/...>
- Badea, Liana, and Mariana Claudia Mungiu-Pupăzan. 2021. "The Economic and Environmental Impact of Bitcoin." *IEEE Access* 9: 48091–104.
- Bailey, Jason. 2018a. "Collective Behind Christie's AI Art Tells All." *Artnome*. <https://www.artnome.com/news/2018/10/13/collective-behind-christies-ai-art-tells-all> (7 April 2023).
- . 2018b. "The AI Art at Christie's Is Not What You Think." *Artnome*. <https://www.artnome.com/news/2018/10/13/the-ai-art-at-christies-is-not-what-you-think> (7 April 2023).
- . 2018c. "The Truth Behind Christie's \$432K AI Art Sale." *Artnome*. <https://www.artnome.com/news/2018/10/13/the-truth-behind-christies-432k-ai-art-sale> (7 April 2023).
- . 2019. "Why Is AI Art Copyright So Complicated?" *Artnome*. <https://www.artnome.com/news/2019/3/27/why-is-ai-art-copyright-so-complicated> (12 April 2023).
- Bak, Per, Kan Chen, and Michael Creutz. 1989. "Self-Organized Criticality in the Game of Life." *Nature* 342(6251): 780–82.

- Banū-Mūsā, Donald Routledge Hill, and Banū-Mūsā. 1979. *The Book of Ingenious Devices: [Muhammad Ibn-Mūsā Ibn-Šākīr]; Kitāb al-Ĥiyāl. By the Banu (Sons of) Mūsā Bin Šākīr. Transl. and Ann. by Donald R[Outledge] Hill.* Dordrecht [usw.]: Reidel.
- Barratt, Daniel, Anna Cabak Rédei, Åse Innes-Ker, and Joost van de Weijer. 2016. “Does the Kuleshov Effect Really Exist? Revisiting a Classic Film Experiment on Facial Expressions and Emotional Contexts.” *Perception* 45(8): 847-74.
- Bateson, Gregory. 2000. *Steps to an Ecology of Mind.* University of Chicago Press. Chicago: University of Chicago Press.
- Bau, David et al. 2018. “GAN Dissection: Visualizing and Understanding Generative Adversarial Networks.” <http://arxiv.org/abs/1811.10597> (8 February 2023).
- Baumgarten, Alexander Gottlieb. 1750. *Aesthetica.* impens. Ioannis Christiani Kleyb.
- Beer, S. 1959. *Cybernetics and Management.* Wiley.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, 610-23. <https://dl.acm.org/doi/10.1145/3442188.3445922> (19 April 2023).
- Benjamin, Walter et al. 2008. *The Work of Art in the Age of Its Technological Reproducibility, and Other Writings on Media.* Cambridge, Mass: Belknap Press of Harvard University Press.
- Bense, Max. 1998. “Einführung in die informationstheoretische Ästhetik: Grundlegung und Anwendung in der Texttheorie.” In *Max Bense: Ästhetik und Texttheorie*, Stuttgart: J.B. Metzler, 251-417. http://link.springer.com/10.1007/978-3-476-03716-9_3 (3 March 2023).
- Berlekamp, Elwyn R., John Horton Conway, and Richard K. Guy. 1985. *Winning Ways for Your Mathematical Plays. 2: Games in Particular.* 3rd print. with corrections. London: Academic Press.
- Bertalanffy, L. von. 1945. “Zu Einer Allgemeinen Systemlehre. Blaetter f.” *Dtsche. Philos.* y 18.
- Bertalanffy, Ludwig von. 2009. *General System Theory: Foundations, Development, Applications.* Rev. ed., 17th paperback print. New York, NY: Braziller.
- Betker, James et al. 2024. “Improving Image Generation with Better Captions.” <https://cdn.openai.com/papers/dall-e-3.pdf>.
- Biletzki, Anat, and Anat Matar. 2021. “Ludwig Wittgenstein.” In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2021/entries/wittgenstein/>.
- Bishop, J. Mark, and Mohammad Majid Al-Rifaie. 2016. “Autopoiesis in Creativity and Art.” In *Proceedings of the 3rd International Symposium on Movement and Computing*, 1-6.
- Bishop, J. Mark, and Mohammad M. al-Rifaie. 2017. “Autopoiesis, Creativity and Dance.” *Connection Science* 29(1): 21-35.
- Boden, M.A. 2018. *Artificial Intelligence: A Very Short Introduction.* OUP Oxford. <https://books.google.es/books?id=mB5pDwAAQBAJ>.
- Borgdorff, Henk. 2006. *2 The Debate on Research in the Arts.* Kunsthøgskolen i Bergen Bergen, Norway.
- “Botto - The Manifesto.” <https://www.botto.com/dashboard/manifesto> (3 February 2023).

- Bourdieu, Pierre. 2002. *Distinction: A Social Critique of the Judgement of Taste*. 11. print. Cambridge, Mass: Harvard Univ. Press.
- Bown, Oliver, and Jon McCormack. 2011. "Creative Agency: A Clearer Goal for Artificial Life in the Arts." In *Advances in Artificial Life. Darwin Meets von Neumann*, Lecture Notes in Computer Science, eds. George Kampis, István Karsai, and Eörs Szathmáry. Berlin, Heidelberg: Springer, 254–61.
- Bradski, Gary R., and Adrian Kaehler. 2011. *Learning OpenCV: Computer Vision with the OpenCV Library*. 1st ed., [Nachdr.]. Beijing: O'Reilly.
- Brembs, Björn. 2010. "Towards a Scientific Concept of Free Will as a Biological Trait: Spontaneous Actions and Decision-Making in Invertebrates." *Proceedings of the Royal Society B: Biological Sciences* 278(1707): 930–39.
- Bringsjord, Selmer, Paul Bello, and David Ferrucci. 2003. "Creativity, the Turing Test, and the (Better) Lovelace Test." In *The Turing Test: The Elusive Standard of Artificial Intelligence*, Studies in Cognitive Systems, ed. James H. Moor. Dordrecht: Springer Netherlands, 215–39. https://doi.org/10.1007/978-94-010-0105-2_12 (26 October 2021).
- Broad, Terence, and Mick Grierson. 2016. *Autoencoding Video Frames*. London: Goldsmiths. Report. https://www.academia.edu/25585807/Autoencoding_Video_Frames (12 April 2023).
- Brock, Andrew, Jeff Donahue, and Karen Simonyan. 2019. "Large Scale GAN Training for High Fidelity Natural Image Synthesis." <http://arxiv.org/abs/1809.11096> (18 November 2022).
- Bromley, A.G. 2000. "Babbage's Analytical Engine Plans 28 and 28a. The Programmer's Interface." *IEEE Annals of the History of Computing* 22(4): 5–19.
- Bromley, Allan G. 1982. "Charles Babbage's Analytical Engine, 1838." *IEEE Annals of the History of Computing* 4(3): 196–217.
- Brown, Tom B. et al. 2020. "Language Models Are Few-Shot Learners." <http://arxiv.org/abs/2005.14165> (31 August 2023).
- Buchanan, Bruce G. 2005. "A (Very) Brief History of Artificial Intelligence." *AI Magazine* 26(4): 53.
- Burnell, Ryan et al. 2023. "Rethink Reporting of Evaluation Results in AI." *Science* 380(6641): 136–38.
- Burnham, Jack. 1968. "Systems Esthetics." *Artforum* 7(1): 30–35.
- . 1974. *Great Western Salt Works: Essays on the Meaning of Post-Formalist Art*. New York: G. Braziller.
- Bush, Vannevar. 1945. "As We May Think." *The Atlantic*. <https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/> (10 October 2022).
- Caldas Vianna, Bruno. 2020. "Generative Art: Between the Nodes of Neuron Networks." *Artnodes* (26). <https://www.raco.cat/index.php/Artnodes/article/view/374003> (9 September 2021).
- . 2021a. "Aiku." <https://github.com/brunovianna/aiku> (28 April 2023).
- . 2021b. "Singular." *Research Catalogue*. <https://www.researchcatalogue.net/view/1408828/1408854> (28 April 2023).
- . 2022. "Knowledge Art: Artwork and Invention." In *10th International Conference on Digital and Interactive Arts, ARTECH 2021*, New York, NY, USA: Association for Computing Machinery, 1–7. <https://doi.org/10.1145/3483529.3483653> (6 February 2023).
- . 2023. "Adapted for Post-Colonialism Piece." <https://github.com/brunovianna/face-classification-postcolonialism> (12 April 2023).

- Caldas Vianna, Bruno. 2023. "Brunovianna/Language-Walks." <https://github.com/brunovianna/language-walks> (15 April 2023).
- Calleja Reina, Rocío. 2023. "La Inteligencia Artificial y Su Derivada En Los Derechos de Propiedad Intelectual En La Cultura: Retos y Amenazas." *Periferica* (24). <https://revistas.uca.es/index.php/periferica/article/view/10720> (6 February 2024).
- Campbell, Murray. 1999. "Knowledge Discovery in Deep Blue." *Communications of the ACM* 42(1): 65–67.
- Cardona, George. 1997. *Pāṇini, His Work and Its Traditions*. 2. rev. & enl. ed. Delhi: Motilal Banarsidass Publishers.
- Carroll, Noel. 1994. "Visual Metaphor." *Aspects of Metaphor*: 189–218.
- Chalmers, David John. 1997. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
- Chollet, François. 2017. "GANism (the Specific Look and Feel of Seemingly GAN-Generated Images) May yet Become a Significant Modern Art Trend." *Twitter*. <https://twitter.com/fchollet/status/885378870848901120> (10 April 2023).
- . 2018. *Deep Learning with Python*. Shelter Island, New York: Manning Publications Co.
- Church, Alonzo. 1936. "An Unsolvable Problem of Elementary Number Theory." *American Journal of Mathematics* 58(2): 345–63.
- Clark, Andy. 2017. "How to Knit Your Own Markov Blanket." In *Philosophy and Predictive Processing*, eds. Thomas Metzinger and Wanja Wiese.
- Clark, David. 1997. "Deep Thoughts on Deep Blue." *IEEE Computer Architecture Letters* 12(04): 31–31.
- Clarke, Bruce, and Mark B. N. Hansen, eds. 2009. *Emergence and Embodiment: New Essays on Second-Order Systems Theory*. Durham: Duke University Press.
- Collier, Bruce, and MacLachlan, James. 1998. *Charles Babbage and the Engines of Perfection*. New York; Oxford: Oxford University Press.
- Copeland, B. Jack, ed. 2006. *Colossus: The Secrets of Bletchley Park's Codebreaking Computers*. Oxford; New York: Oxford University Press.
- . 2020. "The Modern History of Computing." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2020/entries/computing-history/> (20 October 2022).
- Copeland, Jack, Jonathan Bowen, Mark Sprevak, and Robin Wilson. 2017. *The Turing Guide*.
- Crevier, Daniel. 1993. *AI: The Tumultuous History of the Search for Artificial Intelligence*. Basic Books.
- Crowson, Katherine et al. 2022. "VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance." <http://arxiv.org/abs/2204.08583> (18 November 2022).
- Dasgupta, Subrata. 2014. *It Began with Babbage: The Genesis of Computer Science*. ed. Subrata Dasgupta. New York: Oxford, University Press.
- Delacroix, Sylvie. 2021. "Computing Machinery, Surprise and Originality." *Philosophy & Technology*. <https://link.springer.com/10.1007/s13347-021-00453-8> (26 October 2021).
- Deleuze, Gilles. 1989. *Cinema. 2: The Time-Image*. London: Athlone.
- Deng, Jia et al. 2009. "ImageNet: A Large-Scale Hierarchical Image Database." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: IEEE, 248–55. <https://ieeexplore.ieee.org/document/5206848/> (31 March 2023).

- Dhariwal, Prafulla, and Alex Nichol. 2021. "Diffusion Models Beat GANs on Image Synthesis." <http://arxiv.org/abs/2105.05233> (17 November 2022).
- Dreher, Thomas. 2005. "Conceptual Art and Software Art: Notations, Algorithms and Codes." In *Lecture Presented*, https://www.academia.edu/download/42280677/Conceptual_Art_and_Software_Art.pdf (28 August 2023).
- . 2020. *History of Computer Art*. Second. Morrisville, North Carolina: Lulu Press, inc.
- Duchamp, Marcel. 1917. "The Richard Mutt Case." *The Blind Man* 2(5): 817–18.
- Eco, Umberto. 1986. *Semiotics and the Philosophy of Language*. Bloomington: Indiana University Press.
- . 2017. "The Ars Magna by Ramon Llull." *Contributions to Science* 12(1). <http://revistes.iec.cat/index.php/CtS/article/view/142865/141755>.
- Edwards, Benj. 2022. "China Bans AI-Generated Media without Watermarks." *Ars Technica*. <https://arstechnica.com/information-technology/2022/12/china-bans-ai-generated-media-without-watermarks/> (24 April 2023).
- "Eidetic Reduction | Philosophy | Britannica." 2023. <https://www.britannica.com/topic/eidetic-reduction> (14 February 2023).
- Elbeshbishi, Sarah. 2022. "Inspiration or Infringement? Songwriters Clashing in Court More Often after 'Blurred Lines.'" *USA TODAY*. <https://www.usatoday.com/story/entertainment/music/2022/04/24/copyright-infringement-cases-increase/7367700001/> (7 August 2023).
- Eliasmith, Chris. 2015. "Building a Behaving Brain." In *The Future of the Brain: Essays by the World's Leading Neuroscientists*, Princeton University Press, 125–39.
- Esser, Patrick, Robin Rombach, and Björn Ommer. 2021. "Taming Transformers for High-Resolution Image Synthesis." <http://arxiv.org/abs/2012.09841> (18 November 2022).
- Fellbaum, Christiane. 2010. "WordNet." In *Theory and Applications of Ontology: Computer Applications*, eds. Roberto Poli, Michael Healy, and Achilles Kameas. Dordrecht: Springer Netherlands, 231–43. https://link.springer.com/10.1007/978-90-481-8847-5_10 (30 March 2023).
- Fisher, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." In *Annals of Eugenics* 7(2): 179–88.
- Flusser, Vilém. 2011. *Into the Universe of Technical Images*. Minneapolis: University of Minnesota Press.
- Franco, Francesca. 2022. "Interview with Vera Molnar." In *The Algorithmic Dimension: Five Artists in Conversation*, Springer Series on Cultural Computing, ed. Francesca Franco. Cham: Springer International Publishing, 83–100. https://doi.org/10.1007/978-3-319-61167-9_4 (26 March 2023).
- Freeth, Tony et al. 2006. "Decoding the Ancient Greek Astronomical Calculator Known as the Antikythera Mechanism." *Nature* 444(7119): 587–91.
- Frege, Gottlob. 1948. "Sense and Reference." *The Philosophical Review* 57(3): 209–30.
- Friston, Karl, James Kilner, and Lee Harrison. 2006. "A Free Energy Principle for the Brain." *Journal of Physiology-Paris* 100(1–3): 70–87.
- Fukushima, Kunihiko. 1975. "Cognitron: A Self-Organizing Multilayered Neural Network." *Biological Cybernetics* 20(3–4): 121–36.
- . 1980. "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position." *Biological Cybernetics* 36(4): 193–202.

- Galanter, Philip. 2003. "What Is Generative Art? Complexity Theory as a Context for Art Theory." In *Milano*, 21.
- . 2016. "Generative Art Theory." *A Companion to Digital Art*: 146–80.
- Gardner, Martin. 1970. "Mathematical Games." *Scientific American* 223(4): 120–23.
- Gehér, Klára Borbla. 2018. "Nicolas Schöffer Archiv und Vermittlung." . PhD Thesis, TU Wien. doi:10.34726/HSS.2018.34984.
- Gebru Timnit [@timnitGebru]. 2023a. "@MaxZiebell @deguerre @Grady_Booch @EMostaque @spawning_ @CommonCrawl Its Not Inspired Work: <https://T.Co/nMgfhbgSAv>." *Twitter*. <https://twitter.com/timnitGebru/status/1630046752316493825> (24 April 2023).
- . 2023b. "My Kind of Computer Scientists. Via @mmitchell_ai Makes Me so Happy to See Work like This. GLAZE: Protecting Artists from Style Mimicry by Text-to-Image Models By Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, Ben Y. Zhao <https://Arxiv.Org/Pdf/2302.04222.Pdf>." *Twitter*. <https://twitter.com/timnitGebru/status/1625323008457048065> (24 April 2023).
- Gershgorn, Dave. 2017. "The Data That Transformed AI Research—and Possibly the World." *Quartz*. <https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/> (7 April 2023).
- Giles, Martin. 2018. "The GANfather: The Man Who's Given Machines the Gift of Imagination." *MIT Technology Review*: 10.
- Gödel, Kurt. 1992. *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*. New York: Dover Publications. https://monoskop.org/images/9/93/Kurt_G%C3%B6del_On_Formally_Undecidable_Propositions_of_Principia_Mathematica_and_Related_Systems_1992.pdf.
- Goldstine, Herman H. 1993. *The Computer from Pascal to von Neumann*. Course Book. Princeton, N.J: Princeton University Press.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, Massachusetts: The MIT Press.
- Goodfellow, Ian J. et al. 2014. "Generative Adversarial Networks." *arXiv:1406.2661 [cs, stat]*. <http://arxiv.org/abs/1406.2661> (22 October 2019).
- "GPT-2: 1.5B Release." 2019. *OpenAI Research*. <https://web.archive.org/web/20230331004642/https://openai.com/research/gpt-2-1-5b-release> (31 August 2023).
- Grayling, A. C. 2001. *Wittgenstein: A Very Short Introduction*. Oxford: Oxford University Press.
- Griewank, Andreas. 2012. "Who Invented the Reverse Mode of Differentiation."
- Gui, Jie et al. 2020. "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications." <http://arxiv.org/abs/2001.06937> (4 April 2023).
- Gullberg, Jan. 1997. *Mathematics: From the Birth of Numbers*. 1st ed. New York: W.W. Norton.
- Gunzenhäuser, Rul. 1962. *Ästhetisches Mass Und Ästhetische Information: Einführung in Die Theorie GD Birkhoffs Und Die Redundanztheorie Ästhetischer Prozesse*. Schnelle.
- Haenlein, Michael, and Andreas Kaplan. 2019. "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence." *California Management Review* 61(4): 5–14.
- Halsall, Francis. 2008. *Systems of Art: Art, History and Systems Theory*. Bern; Oxford: Peter Lang.
- Hannay, Chris. 2018. "Artist Faces Lawsuit over Computer System That Creates Randomly Generated Images." *The Globe and Mail*. <https://www.theglobeandmail.com>.

- com/arts/art-and-architecture/article-artist-faces-lawsuit-over-computer-system-that-creates-randomly/ (21 April 2023).
- Haraway, Donna Jeanne. 1991. *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York: Routledge.
- Haskins, Casey. 1990. "Kant, Autonomy, and Art for Art's Sake." *The Journal of Aesthetics and Art Criticism* 48(3): 235.
- Haugeland, John. 1986. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.
- Hayles, N. Katherine. 1999. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago, Ill: University of Chicago Press.
- . 2017. *Unthought: The Power of the Cognitive Nonconscious*. Chicago; London: The University of Chicago Press.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Deep Residual Learning for Image Recognition." <http://arxiv.org/abs/1512.03385> (18 May 2022).
- Heaven, Will Douglas. 2023. "GPT-4 Is Bigger and Better than ChatGPT—but OpenAI Won't Say Why." *MIT Technology Review*. <https://web.archive.org/web/20230317224201/https://www.technologyreview.com/2023/03/14/1069823/gpt-4-is-bigger-and-better-chatgpt-openai/> (31 August 2023).
- Heng-Tze Cheng. 2022. "Google AI Blog: LaMDA: Towards Safe, Grounded, and High-Quality Dialog Models for Everything." <https://web.archive.org/web/20220325014118/https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html> (19 October 2022).
- Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33(2–3): 61–83.
- Hilbert, David. 1902. "Mathematical Problems." *Bulletin of the American Mathematical Society* 8(10): 437–79.
- Hinton, Geoffrey E., and Sam Roweis. 2002. "Stochastic Neighbor Embedding." *Advances in Neural Information Processing Systems* 15.
- Hodges, Andrew. 2014. *Alan Turing: The Enigma: The Book That Inspired the Film The Imitation Game - Updated Edition*.
- Hofstadter, Douglas R. 2000. *Gödel, Escher, Bach: An Eternal Golden Braid*. 20th Anniversary ed. London: Penguin.
- Hubel, D. H., and T. N. Wiesel. 1959. "Receptive Fields of Single Neurones in the Cat's Striate Cortex." *The Journal of Physiology* 148(3): 574–91.
- . 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *The Journal of Physiology* 160(1): 106–54.
- "How Ian Goodfellow Invented GANs." 2020. *How Ian Goodfellow Invented GANs*. <https://www.deeplearning.ai/the-batch/ian-goodfellow-a-man-a-plan-a-gan/> (3 April 2023).
- Hui, Yuk. 2021. *Art and Cosmotronics*. U of Minnesota Press.
- Humphries, Mark. 2019. "Your Cortex Contains 17 Billion Computers." <https://web.archive.org/web/20190808105156/https://medium.com/the-spike/your-cortex-contains-17-billion-computers-9034e42d34f2> (21 February 2023).
- Husbands, Philip, and Owen Holland. 2008. "The Ratio Club: A Hub of British Cybernetics." *The Mechanical Mind in History*: 91–148.
- Hutchins, Edwin. 2006. *Cognition in the Wild*. 8. pr. Cambridge, Mass.: MIT Press.
- Ilachinski, Andrew. 2002. *Cellular Automata: A Discrete Universe*. Reprinted. Singapore: World Scientific.

- “Introducing OpenAI.” 2015. *OpenAI Blog*. <https://web.archive.org/web/20190224031626/https://blog.openai.com/introducing-openai/> (31 August 2023).
- Irvine, Andrew David. 2022. “Bertrand Russell.” In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2022/entries/russell/> (24 February 2023).
- Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2018. “Image-to-Image Translation with Conditional Adversarial Networks.” <http://arxiv.org/abs/1611.07004> (7 April 2023).
- Jonas, Eric, and Konrad Paul Kording. 2017. “Could a Neuroscientist Understand a Microprocessor?” *PLOS Computational Biology* 13(1): e1005268.
- Jones, Jonathan. 2018. “A Portrait Created by AI Just Sold for \$432,000. But Is It Really Art?” *The Guardian*. <https://www.theguardian.com/artanddesign/shortcuts/2018/oct/26/call-that-art-can-a-computer-be-a-painter> (15 October 2019).
- Jorgensen, Timothy J. 2022. “Is the Human Brain a Biological Computer?” <https://press.princeton.edu/ideas/is-the-human-brain-a-biological-computer> (21 February 2023).
- Judovitz, Dalia, and Marcel Duchamp. 2010. *Drawing on Art: Duchamp and Company*. Minneapolis: University of Minnesota Press.
- Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. 1st ed. New York: Farrar, Straus and Giroux.
- Kant, Immanuel. 2000. *Critique of the Power of Judgment*. 1st ed. Paul Guyer. Cambridge University Press. <https://www.cambridge.org/core/product/identifier/9780511804656/type/book> (17 April 2023).
- Karras, Tero et al. 2021. “Alias-Free Generative Adversarial Networks.” In *Proc. NeurIPS*, <https://arxiv.org/abs/2106.12423>.
- Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. “Progressive Growing of GANs for Improved Quality, Stability, and Variation.” <http://arxiv.org/abs/1710.10196> (7 April 2023).
- Karras, Tero, Samuli Laine, and Timo Aila. 2019. “A Style-Based Generator Architecture for Generative Adversarial Networks.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–10.
- Kaufman, James C., and Ronald A. Beghetto. 2009. “Beyond Big and Little: The Four C Model of Creativity.” *Review of General Psychology* 13(1): 1–12.
- Kautz, Henry A. 2022. “The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture.” *AI Magazine* 43(1): 105–25.
- Keenan, Bernard. 2022. “Niklas Luhmann: What Is Autopoiesis?” *Critical Legal Thinking*. <https://criticallegalthinking.com/2022/01/10/niklas-luhmann-what-is-autopoiesis/> (7 June 2023).
- Kekola, Miika. 2020. “AI, Author and Copyright.” Lapland University. <https://urn.fi/URN:NBN:fi-fe2020061644627>.
- Kelly, Sean. 2019. “What Computers Can’t Create: Why Creativity Is, and Always Will Be, a Human Endeavor. - Free Online Library.” MIT Technology Review. <https://www.thefreelibrary.com/What+computers+can%27t+create%3A+Why+creativity+is%2C+and+always+will+be%2C+a...-a0578441401> (27 October 2021).
- Kemper, Jonathan. 2023. “Artists Remove 80 Million Images from Stable Diffusion 3 Training Data.” *THE DECODER*. <https://the-decoder.com/artists-remove-80-million-images-from-stable-diffusion-3-training-data/> (24 April 2023).
- Knuth, Donald E. 1972. “Ancient Babylonian Algorithms.” *Communications of the ACM* 15(7): 671–77.

- Koetsier, Teun. 2001. "On the Prehistory of Programmable Machines: Musical Automata, Looms, Calculators." *Mechanism and Machine Theory* 36(5): 589–603.
- Kogan, Gene. 2021. "Artist in the Cloud." *Medium*. <https://medium.com/@genekogan/artist-in-the-cloud-8384824a75c7> (2 February 2022).
- Koh, Jing Yu, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. "Text-to-Image Generation Grounded by Fine-Grained User Attention." <http://arxiv.org/abs/2011.03775> (10 April 2023).
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2017. "ImageNet Classification with Deep Convolutional Neural Networks." *Communications of the ACM* 60(6): 84–90.
- Kuleshov, Lev Vladimirovich. 1974. *Kuleshov on Film: Writings*. Univ of California Press.
- Laion. "LAION-Aesthetics | LAION." <https://laion.ai/blog/laion-aesthetics> (3 March 2023).
- Larsen, Anders Boesen Lindbo, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. "Autoencoding beyond Pixels Using a Learned Similarity Metric." In *Proceedings of the 33rd International Conference on Machine Learning*, PMLR, 1558–66. <https://proceedings.mlr.press/v48/larsen16.html> (12 April 2023).
- Lebel, Robert et al. 2021. *Marcel Duchamp*. Facsimile edition. Zürich: Hauser & Wirth Publishers.
- Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 86(11): 2278–2324.
- LeCun, Yann. 1985. "Une Procedure d'apprentissage Ponr Réseau a Seuil Asymetrique." *Proceedings of Cognitiva* 85: 599–604.
- . 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1(4): 541–51.
- Lee, Rosemary. 2018. "The Limits of Algorithmic Perception: Technological Umwelt." <https://scienceopen.com/document?vid=684d2f5f-eb85-4afc-b730-9a2d6208f60c> (26 August 2020).
- Lehman, Joel, and Kenneth O. Stanley. 2011. "Abandoning Objectives: Evolution Through the Search for Novelty Alone." *Evolutionary Computation* 19(2): 189–223.
- Leibniz, Gottfried Wilhelm, Freiherr von, 1646-1716., and Philip P. Wiener. 1951. *Leibniz: Selections*. New York: Scribner.
- Lenat, Doug, Mayank Prakash, and Mary Shepherd. 1986. "CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks." *AI Magazine* 6(4): 65–85.
- Leonard, Molly, Cheri Shakiban, and La Yupana. "The Incan Abacus: A Curious Counting Device."
- Leventov, Roman. 2022. "Properties of Current AIs and Some Predictions of the Evolution of AI from the Perspective of Scale-Free Theories of Agency and Regulative Development." *Less Wrong*. <https://www.lesswrong.com/posts/oSPHmfnMGgGrpe7ib/properties-of-current-ais-and-some-predictions-of-the> (24 January 2023).
- Levin, Michael. 2022. "Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds." *Frontiers in Systems Neuroscience* 16. <https://www.frontiersin.org/articles/10.3389/fnsys.2022.768201> (27 January 2023).
- Li, Wei, Melvin Gauci, and Roderich Gross. 2013. "A Coevolutionary Approach to Learn Animal Behavior through Controlled Interaction." In *Proceedings of the 15th Annual*

- Conference on Genetic and Evolutionary Computation*, Amsterdam The Netherlands: ACM, 223–30. <https://dl.acm.org/doi/10.1145/2463372.2465801> (4 April 2023).
- Li, Wenbo et al. 2019. “Object-Driven Text-to-Image Synthesis via Adversarial Training.” <http://arxiv.org/abs/1902.10740> (10 April 2023).
- Lin, Tsung-Yi et al. 2015. “Microsoft COCO: Common Objects in Context.” <http://arxiv.org/abs/1405.0312> (10 April 2023).
- Lindberg, Van. 2023. “A Mixed Decision from the US Copyright Office.” *Process Mechanics*. <https://www.processmechanics.com/2023/02/22/a-mixed-decision-from-the-us-copyright-office/> (24 April 2023).
- Linsky, Bernard, and Andrew David Irvine. 2022. “Principia Mathematica.” In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2022/entries/principia-mathematica/>.
- Liu, Ziwei, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. “Deep Learning Face Attributes in the Wild.” <http://arxiv.org/abs/1411.7766> (8 April 2023).
- llyasviel. 2023. “AdverseCleaner.” <https://github.com/llyasviel/AdverseCleaner> (24 April 2023).
- Lloyd, S. 2006. “Least Squares Quantization in PCM.” *IEEE Transactions on Information Theory* 28(2): 129–37.
- Lovelace, Ada King, and Betty A. Toole. 1998. *Ada, the Enchantress of Numbers: A Selection from the Letters of Lord Byron’s Daughter and Her Description of the First Computer*. 2nd ed. Mill Valley, Calif., Sausalito, CA: Strawberry Press; Orders to Critical Connection.
- Lovelace, Augusta Ada. 1843. “Notes by AAL [Augusta Ada Lovelace].” *Taylor’s Scientific Memoirs*: 666–731.
- Luccioni, Alexandra Sasha, Sylvain Viguier, and Anne-Laure Ligozat. 2022. “Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model.” <http://arxiv.org/abs/2211.02001> (1 September 2023).
- Luhmann, Niklas. 1995. *Social Systems*. Stanford University Press.
- . 2000. *Art as a Social System*. California (Calif.): Stanford University Press.
- Luhmann, Niklas, Klaus A. Ziegert, and Fatima Kastner. 2004. *Law as a Social System*. Oxford; New York: Oxford University Press.
- Lynch, Garrett. 2018. “The Transformative Nature of Networks within Contemporary Art Practice.” London South Bank University. <https://hcommons.org/deposits/item/hc:24123/>.
- Malafouris, Lambros. 2008. “At the Potter’s Wheel: An Argument for Material Agency.” In *Material Agency*, eds. Carl Knappett and Lambros Malafouris. Boston, MA: Springer US, 19–36. http://link.springer.com/10.1007/978-0-387-74711-8_2 (1 February 2024).
- Manovich, Lev. 2019. “Defining AI Arts: Three Proposals.” <http://manovich.net/content/04-projects/107-defining-ai-arts-three-proposals/> (20 April 2023).
- Mansimov, Elman, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2016. “Generating Images from Captions with Attention.” <http://arxiv.org/abs/1511.02793> (17 November 2022).
- Marcus, Gary. 2015. “The Computational Brain.” In *The Future of the Brain*, eds. Gary Marcus and Jeremy Freeman. Princeton University Press, 205–16. <https://www.degruyter.com/document/doi/10.1515/9781400851935-025/html> (16 January 2023).
- Marcus, Gary F., and Jeremy Andrew Freeman, eds. 2015. *The Future of the Brain: Essays by the World’s Leading Neuroscientists*. Princeton: Princeton University Press.

- Marsaglia, George. 1972. "Choosing a Point from the Surface of a Sphere." *The Annals of Mathematical Statistics* 43(2): 645–46.
- Martin, Marcel. 1985. *Le Langage Cinématographique*. 4e éd. rev. et augm. Paris: Cerf.
- Maturana, Humberto R., and Francisco J. Varela. 1980. *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht, Netherlands; Boston: D. Reidel Pub. Co.
- McCarthy, John. 1960. *Programs with Common Sense*. RLE and MIT Computation Center, Cambridge, MA, USA.
- McCorduck, Pamela. 2004. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. 25th anniversary update. Natick, Mass: A.K. Peters.
- McCulloch, Warren S., and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics* 5(4): 115–33.
- McEwan, John D. 1987. "Anarchism and the Cybernetics of Self-Organising Systems." In *A Decade of Anarchy, 1961-1970: Selections from the Monthly Journal Anarchy*, Freedom Press.
- McGinn, Colin. 1999. *The Mysterious Flame: Conscious Minds in a Material World*. 1st ed. New York: Basic Books.
- Minsky, Marvin. 1974. A Framework for Representing Knowledge. MIT, Cambridge.
- Mitchell, Melanie. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. Penguin UK.
- Moles, Abraham A. 1966. *Information Theory and Esthetic Perception*. Urbana: University of Illinois Press.
- Molnar, Vera. 1975. "Toward Aesthetic Guidelines for Paintings with the Aid of a Computer." *Leonardo (Oxford)* 8(3): 185–89.
- Moravec, Hans. 1995. *Mind Children: The Future of Robot and Human Intelligence*. 4th print. Cambridge: Harvard Univ. Press. <https://archive.org/details/mindchildren00hans/page/n3/mode/2up>.
- Moscovich, Viviana Ruth. 2016. *El Khipu y La Yupana: Administración y Contabilidad En El Imperio Inca*. Arequipa, Peru: Ediciones El Lector.
- Moss, Aaron. 2023. "Artists Attack AI: Why The New Lawsuit Goes Too Far." *Copyright Lately*. <https://copyrightlately.com/artists-copyright-infringement-lawsuit-ai-art-tools/> (26 January 2023).
- Nahum, Andrew. 2017. "Automata, Androids and Life." *Robots: The 500-year quest to make machines human' by Ben Russell*: 50–67.
- Nakamoto, Satoshi. 2008. "Bitcoin: A Peer-to-Peer Electronic Cash System." *Decentralized Business Review*: 21260.
- Nake. 2012. "Information Aesthetics: An Heroic Experiment." *Journal of Mathematics and The Arts* 6: 65–75.
- Nake, F. 1968. "Erzeugung ästhetischer Objekte mit Rechenanlagen." In *Nicht-numerische Informationsverarbeitung*, ed. Rul Gunzenhäuser. Vienna: Springer Vienna, 456–72. http://link.springer.com/10.1007/978-3-7091-8202-4_22 (27 March 2023).
- Närhinen, Tuula. 2022. "Do It Yourself, Rain! Dabbling Drops, Splashes, and Waves: Experiments in Art and Science." *Leonardo* 55(6): 627–34.
- Nees, G. 1969. *Generative Computergraphik*. Siemens.
- Neumann, John Von. 1944. *Theory Of Games And Economic Behavior*. <http://archive.org/details/in.ernet.dli.2015.215284> (10 October 2022).
- Newborn, Monty. 2003. *Deep Blue*. New York, NY: Springer New York. <http://link.springer.com/10.1007/978-0-387-21790-1> (4 November 2022).

- Nichol, Alex. 2021. "Openai/Glide-Text2im: GLIDE: A Diffusion-Based Text-Conditional Image Synthesis Model." <https://github.com/openai/glide-text2im> (11 April 2023).
- . 2022. "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models." <http://arxiv.org/abs/2112.10741> (21 November 2022).
- . 2023. "Improved-Diffusion." <https://github.com/openai/improved-diffusion> (11 April 2023).
- Nichol, Alex, and Prafulla Dhariwal. 2021. "Improved Denoising Diffusion Probabilistic Models." <http://arxiv.org/abs/2102.09672> (17 November 2022).
- Niemitalo, Olli. 2010. "Ideas." *Iki.fi*. <http://yehar.com/blog/?p=167>.
- Niles, Richard. 2014. *The Invisible Artist: Arrangers in Popular Music (1950-2000)*. Createspace Independent Publishing Platform.
- Nilsback, Maria-Elena, and Andrew Zisserman. 2008. "Automated Flower Classification over a Large Number of Classes." In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, IEEE, 722–29.
- Noll, A. Michael. 1994. "The Beginnings of Computer Art in the United States: A Memoir." *Leonardo* 27(1): 39.
- Nunes, Rodrigo. 2017. "Multidão e Organização: Plano Ou Sujeito." *Negri no trópico* 23(26(14): 123–40.
- "NVlabs/Ffhq-Dataset." 2023. <https://github.com/NVlabs/ffhq-dataset> (8 April 2023).
- "Nyaya | Indian Philosophy | Britannica." <https://www.britannica.com/topic/Nyaya> (4 May 2022).
- Olazaran, Mikel. 1996. "A Sociological Study of the Official History of the Perceptrons Controversy." *Social Studies of Science* 26(3): 611–59.
- "Openai/Gpt-2." 2019. <https://github.com/openai/gpt-2> (28 April 2023).
- Ortiz, Karla. 2022. "Why AI Models Are Not Inspired like Humans." *KOrtiz Blog*. <https://www.kortizblog.com/blog/why-ai-models-are-not-inspired-like-humans> (24 April 2023).
- . 2022. "1/ You Know How I Know AI Companies like Stability AI, Knowingly and Purposefully Took and Used Visual Artists and General Public Copyrighted Data and Private Data without a Care or Concern (aside from Seeing Data on Where They Scraped Their Data From.?" *Twitter*. <https://twitter.com/kortizart/status/1576092602524323840> (24 April 2023).
- Parikka, Jussi. 2023. "Operational Images: Between Light and Data." *e-flux journal* (133). <https://www.e-flux.com/journal/133/515812/operational-images-between-light-and-data/> (6 January 2023).
- Pask, Gordon. 1976. "Conversation Theory." *Applications in Education and Epistemology*.
- Păun, Gheorghe. 2019. "Some Wonders of a Bio-Computer-Scientist." *Annals of West University of Timisoara - Mathematics and Computer Science* 57(1): 97–121.
- Peano, Giuseppe. 1889. *Arithmetices Principia: Nova Methodo Exposita*. Fratres Bocca.
- Peirce, C. S. 1881. "On the Logic of Number." *American Journal of Mathematics* 4(1): 85–95.
- Peirce, Charles S. 1865. "Five Hundred and Eightieth Meeting. March 12, 1867. Adjourned Statute Meeting; On an Improvement in Boole's Calculus of Logic." In *Proceedings of the American Academy of Arts and Sciences*, JSTOR, 249–61.
- Peirce, Charles S. 2019. "Abductive Reasoning and the Genesis of New Ideas." *The Creativity Reader*: 157.
- Perrigo, Billy. 2023. "Exclusive: The \$2 Per Hour Workers Who Made ChatGPT Safer." *Time*. <https://time.com/6247678/openai-chatgpt-kenya-workers/> (6 March 2023).

- Petőfi, János S. 1976. "A Frame for FRAMES." *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*. <https://www.scinapse.io/papers/2127662966> (7 March 2023).
- Pinker, Steven. 2010. *The Language Instinct: How the Mind Creates Language*. New York: Harper Perennial.
- Prabhu, Vinay Uday, and Abeba Birhane. 2020. "Large Image Datasets: A Pyrrhic Win for Computer Vision?" *arXiv preprint arXiv:2006.16923*.
- Pressman, John David. 2023a. "JD-P/Simulacra-Aesthetic-Captions." <https://github.com/JD-P/simulacra-aesthetic-captions#participants-are-weird> (6 March 2023).
- . 2023b. "Simulacra Aesthetic Captions Bot." <https://github.com/JD-P/simulacrabot> (6 March 2023).
- Pressman, John David [@jd_pressman]. 2022. "A Fundamental Problem with RLHF Is That the Process Doesn't Extract Human Values, but Human Values Conditional on the Models Biases. As a Straightforward Example: 🗿." *Twitter*. https://twitter.com/jd_pressman/status/1535700144011612161 (6 March 2023).
- Radford, Alec et al. 2021. "Learning Transferable Visual Models From Natural Language Supervision." *arXiv:2103.00020 [cs]*. <http://arxiv.org/abs/2103.00020> (2 February 2022).
- Radford, Alec, Luke Metz, and Soumith Chintala. 2016. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks." *arXiv:1511.06434 [cs]*. <http://arxiv.org/abs/1511.06434> (22 October 2019).
- Rainer, Anton et al. 2023. "(S+) Wie der Hype um künstliche Intelligenz einen globalen Machtkampf ausgelöst hat." *Der Spiegel*. <https://www.spiegel.de/netzwelt/web/chatgpt-und-co-wie-der-hype-um-ki-einen-globalen-machtkampf-ausgeloesht-hat-a-a41ce586-95a2-4516-b3bd-6348d9372781> (6 March 2023).
- Ramesh, Aditya et al. 2021. "Zero-Shot Text-to-Image Generation." <http://arxiv.org/abs/2102.12092> (17 November 2022).
- . 2022. "Hierarchical Text-Conditional Image Generation with CLIP Latents." <http://arxiv.org/abs/2204.06125> (21 November 2022).
- . 2023. "DALL-E." <https://github.com/openai/DALL-E> (10 April 2023).
- Ramponi, Marco. 2023. "The Full Story of Large Language Models and RLHF." *AssemblyAI News, Tutorials, AI Research*. <https://www.assemblyai.com/blog/the-full-story-of-large-language-models-and-rlhf/> (1 September 2023).
- Ramstead, Maxwell James Désormeau, Paul Benjamin Badcock, and Karl John Friston. 2018. "Answering Schrödinger's Question: A Free-Energy Formulation." *Physics of Life Reviews* 24: 1–16.
- Randell, Brian. 1982. "From Analytical Engine to Electronic Digital Computer: The Contributions of Ludgate, Torres, and Bush." *Annals of the History of Computing* 4(4): 327–41.
- Raviv, Shaun. 2018. "The Genius Neuroscientist Who Might Hold the Key to True AI | WIRED." <https://web.archive.org/web/20181113154312/https://www.wired.com/story/karl-friston-free-energy-principle-artificial-intelligence/> (26 January 2023).
- Reed, Scott et al. 2016. "Generative Adversarial Text to Image Synthesis." <http://arxiv.org/abs/1605.05396> (17 November 2022).
- Reichardt, Jasia. 1969. *Cybernetic Serendipity: The Computer and the Arts*. Praeger.
- Rivers Have Wings [@RiversHaveWings]. 2021. "@SeanPedersen96 Encode Style with CLIP, Encode Content with VQGAN. in a Loop, Decode Current VQGAN Latent, Encode It with CLIP, Compare Result with Style, Obtain Gradient for VQGAN

- Latent, Update VQGAN Latent." *Twitter*. <https://twitter.com/RiversHaveWings/status/1382830131828334594> (11 April 2023).
- Roberts, Veronica. 2012. "‘Like a Musical Score’: Variability and Multiplicity in Sol LeWitt’s 1970s Wall Drawings." *Master Drawings* 50(2): 193–210.
- Romano, Aja. 2016. "A Guy Trained a Machine to ‘Watch’ Blade Runner. Then Things Got Seriously Sci-Fi." *Vox*. <https://web.archive.org/web/20160602055152/https://www.vox.com/2016/6/1/11787262/blade-runner-neural-network-encoding> (12 April 2023).
- Rombach, Robin et al. 2022. "High-Resolution Image Synthesis with Latent Diffusion Models." <http://arxiv.org/abs/2112.10752> (17 November 2022).
- Rombach, Robin, and Patrick Esser. 2022. "StableDiffusion 1.4 Model Card." <https://huggingface.co/CompVis/stable-diffusion-v1-4> (1 September 2023).
- Rosen, Robert. 2012. *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*. 2nd ed. New York: Springer.
- Rosenblatt, Frank. 1961. *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms*. Cornell Aeronautical Lab Inc Buffalo NY.
- Rosenblueth, Arturo. 1943. "Behavior, Purpose and Teleology." *Philosophy of science* 10(1): 18–24.
- Ruiten, Schelte van, and Mick Wilson, eds. 2013. *SHARE: Handbook for Artistic Research Education*. Amsterdam: ELIA.
- Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323(6088): 533–36.
- Russakovsky, Olga et al. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision (IJCV)* 115(3): 211–52.
- Russell, Bertrand. 1905. "II.—ON DENOTING." *Mind* XIV(4): 479–93.
- Russell, Stuart J., and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach*. Fourth edition. Hoboken: Pearson.
- Saharia, Chitwan et al. 2022. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." <http://arxiv.org/abs/2205.11487> (21 November 2022).
- Sam, Dolgoff, ed. 1979. "The Relevance of Anarchism to Modern Society." In *Contemporary Anarchism*, Routledge.
- Samuel, Arthur L. 1960. "Some Moral and Technical Consequences of Automation—A Refutation." *Science* 132(3429): 741–42.
- . 2000. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 44(1.2): 206–26.
- Schmidhuber, Juergen. 2015. "Deep Learning." *Scholarpedia* 10(11): 32832.
- Scholz, Barbara C., Francis Jeffrey Pelletier, Geoffrey K. Pullum, and Ryan Nefdt. 2022. "Philosophy of Linguistics." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2022/entries/linguistics/whorfianism.html>.
- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and brain sciences* 3(3): 417–24.
- Secrean, Jimmy et al. 2011. "Picbreeder: A Case Study in Collaborative Evolutionary Exploration of Design Space." *Evolutionary computation* 19(3): 373–403.
- Sevänen, Erkki. 2001. "Art as an Autopoietic Sub-System of Modern Society: A Critical Analysis of the Concepts of Art and Autopoietic Systems in Luhmann’s Late Production." *Theory, Culture & Society* 18(1): 75–103.

- Shan, Shawn, Jenna Cryan, et al. 2023. "GLAZE: Protecting Artists from Style Mimicry by Text-to-Image Models." <http://arxiv.org/abs/2302.04222> (24 April 2023).
- Shan, Shawn, Wenxin Ding, et al. 2023. "Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models." <http://arxiv.org/abs/2310.13828> (7 February 2024).
- Shanken, Edward A. 2002a. "12. Cybernetics and Art: Cultural Convergence in the 1960s." In *From Energy to Information*, eds. Bruce Clarke and Linda Dalrymple Henderson. Stanford University Press, 255–78. <https://www.degruyter.com/document/doi/10.1515/9781503619470-021/html> (23 March 2023).
- . 2002b. "Art in the Information Age: Technology and Conceptual Art." *Leonardo* 35(4): 433–38.
- Shannon, Claude Elwood. 1948. "A Mathematical Theory of Communication." *The Bell system technical journal* 27(3): 379–423.
- Simonyan, Karen, and Andrew Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." <http://arxiv.org/abs/1409.1556> (3 April 2023).
- Smith, Brian Cantwell. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: The MIT Press.
- Smith, David Woodruff. 2018. "Phenomenology." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2018/entries/phenomenology/>.
- Sohl-Dickstein, Jascha, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics." In *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, 2256–65. <https://proceedings.mlr.press/v37/sohl-dickstein15.html> (17 November 2022).
- Standage, Tom. 2003. *The Turk: The Life and Times of the Famous Eighteenth-Century Chess-Playing Machine*. Berkley Publishing Group.
- Sternberg, Robert J., ed. 2000. *Handbook of Intelligence*. Cambridge; New York: Cambridge University Press.
- Strawson, P. F. 1950. "On Referring." *Mind* 59(235): 320–44.
- Strickland, Eliza. 2021. "The Turbulent Past and Uncertain Future of AI: Is There a Way out of AI's Boom-and-Bust Cycle?" *IEEE Spectrum* 58(10): 26–31.
- Sutela, Jenna, ed. 2017. *Orgs: From Slime Mold to Silicon Valley and Beyond*. Helsinki: Garret Publications.
- Swann, Thomas. 2021. *Anarchist Cybernetics: Control and Communication in Radical Politics*. Bristol: Bristol University Press.
- Szegedy, Christian et al. 2014. "Going Deeper with Convolutions." <http://arxiv.org/abs/1409.4842> (12 April 2023).
- Table ronde Infinite Skulls*. 2019. <https://vimeo.com/325843365> (4 August 2023).
- "The Theater: Battle of the Nihilists." 1957. *Time Magazine*. <https://web.archive.org/web/20100525212937/http://www.time.com/time/magazine/article/0,9171,809373,00.html> (23 March 2023).
- Tiku, Nitasha. 2022. "The Google Engineer Who Thinks the Company's AI Has Come to Life." *Washington Post*. <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/> (25 January 2023).
- Tipping, Michael E., and Christopher M. Bishop. 1999. "Mixtures of Probabilistic Principal Component Analyzers." *Neural computation* 11(2): 443–82.
- Torres, Emile P. 2021. "Against Longtermism." *Aeon Essays*. <https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo>.
- Turing, A. M. 1937. "On Computable Numbers, with an Application to the

- Entscheidungsproblem." *Proceedings of the London Mathematical Society* s2-42(1): 230–65.
- Turing, Alan M. 2009. "Computing Machinery and Intelligence." In *Parsing the Turing Test*, Springer, 23–65.
- Tzara, Tristan. 1920. "Dada Manifesto on Feeble Love and Bitter Love." *Seven Dada Manifestoes and Lampisteries*: 31–48.
- USCO. 2021. "U.S. Copyright Office, Compendium of U.S. Copyright Office Practices § 313.2 (3d Ed. 2021)." <https://www.copyright.gov/comp3/>.
- Varela, Francisco J., Evan Thompson, and Eleanor Rosch. 2016. *The Embodied Mind: Cognitive Science and Human Experience*. revised edition. Cambridge, Massachusetts; London England: MIT Press.
- Vaswani, Ashish et al. 2017. "Attention Is All You Need." *arXiv:1706.03762 [cs]*. <http://arxiv.org/abs/1706.03762> (1 February 2022).
- Vézina, Brigitte, and Sarah H. Pearson. 2021. "Should CC-Licensed Content Be Used to Train AI? It Depends." *Creative Commons*. <https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/> (24 April 2023).
- Von Bertalanffy, Ludwig. 1950. "An Outline of General System Theory." *British Journal for the Philosophy of Science*.
- Von Foerster, H., ed. 1969. *Purposive Systems: Proceedings of the First Annual Symposium of the American Society for Cybernetics*. Spartan Books. https://books.google.es/books?id=r_RBAAAAIAAJ.
- Von Foerster, Heinz. 2003. "Ethics and Second-Order Cybernetics." *Understanding Understanding: Essays on Cybernetics and Cognition*: 287–304.
- Von Neumann, John. 2017. "The General and Logical Theory of Automata." In *Systems Research for Behavioral Sciences* *systems Research*, Routledge, 97–107.
- Voronoi, Georges. 1908. "Nouvelles Applications Des Paramètres Continus à La Théorie Des Formes Quadratiques. Premier Mémoire. Sur Quelques Propriétés Des Formes Quadratiques Positives Parfaites." *Journal für die reine und angewandte Mathematik (Crelles Journal)* 1908(133): 97–102.
- Waelder, Pau. 2020. "AI CIAC MTL | No 03 - BEYOND GANISM." *CIAC | Centre international d'art contemporain de Montréal*. <http://ciac.ca/en/ai-ciac-ntl-03-01/> (10 April 2023).
- Walsh, Kit. 2023. "How We Think About Copyright and AI Art." *Electronic Frontier Foundation*. <https://www.eff.org/deeplinks/2023/04/how-we-think-about-copyright-and-ai-art-0> (24 April 2023).
- Wang, Xintao et al. 2018. "Esrgan: Enhanced Super-Resolution Generative Adversarial Networks." In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.
- Wang, Xintao, Liangbin Xie, Chao Dong, and Ying Shan. 2021. "Real-Esrgan: Training Real-World Blind Super-Resolution with Pure Synthetic Data." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1905–14.
- Ward, Colin. 1988. *Anarchy in Action*. London: Freedom Press.
- Welinder, Peter et al. 2011. "The Caltech-Ucsd Birds-200-2011 Dataset."
- Wiener, Norbert. 1948. *Cybernetics: Or Control and Communication in the Animal and the Machine*. 1965 Ed. MIT Press.
- . 1960. "Some Moral and Technical Consequences of Automation." *Science* 131(3410): 1355–58.

- Willman, Marshall. 2021. "Logic and Language in Early Chinese Philosophy." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2021/entries/chinese-logic-language/>.
- Wilson, Stephen. 2002. *Information Arts: Intersections of Art, Science, and Technology*. Cambridge, Mass: MIT Press.
- Wittgenstein, Ludwig, and G. E. M. Anscombe. 2003. *Philosophical Investigations: The German Text, with a Revised English Translation*. 3rd ed. Malden, MA: Blackwell Pub.
- Won, Donghyeon. 2023. "Gender and Race Classification with Face Images." <https://github.com/wondonghyeon/face-classification> (8 April 2023).
- Wüst, Karl, and Arthur Gervais. 2018. "Do You Need a Blockchain." <https://www.scinapse.io> (27 January 2022).
- Xagoraris, Zafrios, and Otto Piene. 1991. "THE AUTOMATON THEATER".
- Xu, Tao et al. 2017. "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks." <http://arxiv.org/abs/1711.10485> (10 April 2023).
- Yang, Kaiyu et al. 2022. "A Study of Face Obfuscation in ImageNet." <http://arxiv.org/abs/2103.06191> (31 March 2023).
- Yudkowsky, Eliezer. 2012. "Constructing Fictional Eugenics (LW Edition) - LessWrong" <https://web.archive.org/web/20230402233449/https://www.lesswrong.com/posts/MdbJXRofWkLpX24FD/constructing-fictional-eugenics-lw-edition> (18 April 2023).
- Zargham, M et al. 2023. "Disambiguating Autonomy: Ceding Control in Favor of Coordination in Cybernetic Organizing." *BlockScience Medium*. <https://medium.com/block-science/disambiguating-autonomy-ca84ac87a0bf>.
- Zeilinger, Martin. 2021a. "Generative Adversarial Copy Machines." *Culture Machine*. <https://culturemachine.net/vol-20-machine-intelligences/generative-adversarial-copy-machines-martin-zeilinger/> (23 September 2021).
- . 2021b. *Tactical Entanglements: AI Art, Creative Agency, and the Limits of Intellectual Property*. Lüneburg: Meson Press.
- Zernay, Regina. 2017. "Casting the First Stone: The Future of Music Copyright Infringement Law after Blurred Lines, Stay with Me, and Uptown Funk." *Chapman Law Review* 20: 177.
- Zhang, Han et al. 2017. "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks." <http://arxiv.org/abs/1612.03242> (10 April 2023).
- . 2018. "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks." <http://arxiv.org/abs/1710.10916> (10 April 2023).
- Zhang, Wei, Jun Tanida, Kazuyoshi Itoh, and Yoshiki Ichioka. 1988. "Shift-Invariant Pattern Recognition Neural Network and Its Optical Architecture." In *Proceedings of Annual Conference of the Japan Society of Applied Physics*, 2147–51.
- Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. 2020. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks." <http://arxiv.org/abs/1703.10593> (7 April 2023).
- Zylinska, J. 2020. *AI Art: Machine Visions and Warped Dreams*. Open Humanities Press. <https://books.google.es/books?id=ZbaozQEACAAJ>.

The poetics of autopoiesis: Visual Arts, Autonomy and Artificial Intelligence
Bruno Caldas Vianna

This publication is the written component of the Doctoral Thesis in Fine Arts for the Doctoral Programme at the Academy of Fine Arts at the University of the Arts Helsinki.

PUBLISHER

Academy of Fine Arts, University of the Arts Helsinki

© University of the Arts Helsinki and writer

GRAPHIC DESIGN

Marjo Malin

COVER

The image is a piece from the Voidism series by Bruno Caldas Vianna.
The color palette was chosen with the help of an AI system.

PRINTING

Hansaprint, 2024

ISBN 978-952-353-459-9 (printed)

ISBN 978-952-353-460-5 (pdf)

The publication is available in the University of the Arts publication archive Taju:
<https://taju.uniarts.fi/>.

This research project was supported by a grant from Suomen Kulttuurirahasto (SKR).

The author is also known as Bruno Caldas Pires, as his full name was changed within the Spanish legal system when he was granted Spanish citizenship.

This research project was sparked by the emergence of generative visual tools based on machine learning in the late 2010s. How much creative autonomy did they afford, compared to prior computational generative devices and human creators? Can we now foresee the birth of an artificial artist, which would be completely self-ruled, independent of humans? Where would the poetics of such a creative entity lie? Informed by frameworks from systems theory and computational cognition, the investigation develops the concept of autopoietic aesthetic arrangement, which highlights a particular poetic mechanism within self-generating artworks. The vertiginous development of creative AI in the most recent years allowed this conceptual toolbox to be tested against new generative technologies, in particular the text-to-image semantic imaging tools.

**UNIARTS
HELSINKI**

X ACADEMY OF FINE ARTS

ISBN 978-952-353-459-9



9 789523 534599