Universitat Oberta de Catalunya

artnodes
E-JOURNAL ON ART, SCIENCE AND TECHNOLOGY

https://artnodes.uoc.edu

ARTICLE

NODE "POSSIBLES"

# The Bartleby Machine: exploring creative disobedience in computers

**Bruno Caldas Vianna**
University of the Arts, Helsinki

Date of submission: December 2022
Accepted in: July 2023
Published in: July 2023

## Recommended citation

Caldas Vianna, Bruno. 2023. «The Bartleby Machine: exploring creative disobedience in computers». In: Pau Alsina & Andrés Burbano (coords.). «Possibles». *Artnodes*, no. 32. UOC. [Accessed: dd/mm/aa]. https://doi.org/10.7238/artnodes.v0i32.409664



## Abstract

The idea of disobedient machines is developed from the perspective of the historical and current developments in artificial intelligence (AI). Disobedience is often used in arts and technology as both a theme and a tool. Beyond that, misbehaviour is presented as one of the skills that is indispensable for natural intelligence. The article doesn't delve into the use of AIs as an assistive tool for creation. Instead, it speculates if AIs will afford the emergence of an independent, autonomous artificial creator. Different approaches to AIs are presented, from symbolism to emergism. The affordances of machine learning models are described, as well as their limitations like the incapacity to generate breakthroughs outside of their training data, their determinism, and the inability to use analogies to solve unseen problems. Other missing human (or biological) skills present in art are emotion, goal-less production, and agency, which is a problem even when human volition is studied. The limits of computational formalism are like the limits in mathematical reasoning – it always requires some external rules, or axioms demonstrated, like Gödel's proof. Hofsdtader's theory of consciousness proposes a way to conciliate the fact that human creativity is also based on closed, fixed biological rules. Finally, it is argued that a machine cannot be creative unless it is also able to misbehave. However, computers must follow a set of instructions or they stop functioning – that is the definition of a Turing machine. Hence, we must face the paradox of wanting well-behaved systems, with the limitations of symbolic machines, while at the same time demanding more autonomous, creative outputs. It is paramount to explore algorithmic misbehaviours that could circumvent this paradox for further development of AIs for the arts and society in general.

**Keywords**

artificial intelligence; art; symbolic AI; connectionism; convolutional neural networks; consciousness; Turing machine; volition; machine disobedience

*La Máquina de Bartleby: explorar la desobediencia creativa en los ordenadores*

***Resumen***

*La idea de máquinas desobedientes se desarrolla desde la perspectiva de los desarrollos históricos y actuales en inteligencia artificial (IA). La desobediencia se utiliza a menudo en arte y tecnología como tema y herramienta. Más allá de eso, la desobediencia se presenta como una de las habilidades indispensables para la inteligencia natural. El artículo no profundiza en el uso de las IAs como herramienta de ayuda para la creación. En su lugar, especula si las IA permitirán la aparición de un creador artificial independiente y autónomo. Se presentan diferentes enfoques de IAs, desde el simbolismo hasta el emergismo. Se describen las ventajas de los modelos de aprendizaje automático, así como sus limitaciones, como la incapacidad de generar avances fuera de sus datos de entrenamiento, su determinismo y la incapacidad de usar analogías para resolver problemas inesperados. Otras habilidades humanas (o biológicas) que faltan, presentes en el arte, son la emoción, la producción sin objetivos y la agencia, lo que es un problema incluso cuando se estudia la voluntad humana. Los límites del formalismo computacional son como los límites del razonamiento matemático: siempre requieren algunas reglas externas, o axiomas probados, como la prueba de Gödel. La teoría de la conciencia de Hofsdtader propone una forma de conciliar el hecho de que la creatividad humana también se basa en reglas biológicas cerradas y fijas. Por último, se argumenta que una máquina no puede ser creativa a menos que también pueda desobedecer. Sin embargo, los ordenadores deben seguir un conjunto de instrucciones o dejan de funcionar, es decir, la definición de una máquina de Turing. Por lo tanto, debemos enfrentarnos a la paradoja de querer sistemas obedientes, con las limitaciones de las máquinas simbólicas, al tiempo que exigimos resultados más autónomos y creativos. Es primordial explorar los comportamientos erróneos algorítmicos que podrían eludir esta paradoja para el desarrollo adicional de IAs para las artes y la sociedad en general.*

***Palabras clave***

*inteligencia artificial; arte; IA simbólica; conexionismo; redes neuronales convolucionales; conciencia; Máquina de Turing; voluntad; desobediencia de la máquina*

## Introduction

The story of science fiction is entangled, from the beginning, with tales about entities invented by humans misbehaving. Mary Shelley's Frankenstein (1818), considered by some as the cornerstone of the genre, depicts a creature that evades the control of the protagonist. Asimov's laws of robotics, which first appeared in a short story published in 1942, assure that a robot that follows such rules will never turn against humans (Asimov 2004).

Yet, this is exactly what happens in *Blade Runner*, the screen adaptation of Philip K. Dick's *Do Androids Dream of Electric Sheep?* In the movie, replicants are androids that have gained consciousness and are fighting for survival while humans hunt them. In such a case, the emergence of a consciousness is a misbehaviour in itself; a theme that appeared even in Pinocchio, where the wooden puppet gains a life of its own, only to start lying and mocking its creator, Gepetto.

We connect rebelliousness to self-awareness so much that even the refusal to do conscription service and other duties is termed "conscientious objection". Disobedience is not a rare subject within the creative fields, and even less among artists who incorporate and discuss technology, as we will see in this article. A known proposition states that the rupture of unwritten rules was fundamental for the development of Western art (Hui 2021, 31). Cubism, the Renaissance and conceptual art would not have appeared if their inventors didn't stand in opposition to the hegemonic culture of their time, while Chinese art, for instance, displays a more continuous type of evolution. Therefore, if machines can't afford to display transgressive behaviour, autonomous computer-made creativity would remain crippled when facing the human (and biological, in general) counterpart. But how does science face the possibility of a man-made transgressive entity?

The field of cybernetics was created in the early years of computational theory with the goal of developing (in machines) and understanding (in animals) the mechanisms of self-control. Even if Norbert Wiener never mentions the term *robot* in the founding book of this science, his interest in feedback systems laid the grounds for the development of robotics (Wiener 1948). Systems theory is a direct descendent of cybernetics and, likewise, it took upon solving the issue of the emergence of consciousness.

The most visible offspring of Cybernetics these days is artificial intelligence. Deep learning techniques yielded impressive results in the

last decade, with widespread adoption in commercial applications and sciences from biology to astronomy; it is also of particular interest to art practitioners and researchers.

It is the tremendous success of AI that prompted some of its main researchers to notice how limited these results were in the face of the main goal of the field, namely the creation of an **artificial general intelligence** (AGI), the singular event of a machine having an intellect comparable to a human being. However, proposing a solution to their lack of disobedience is particularly difficult, in the face of the way computers are built - Turing machines are nothing more than instruction-following devices.

The goal of this text is to bring the subject of machine disobedience to the light of some existent artificial intelligence theories and to do so under the perspective of artistic practice. The question here is not about the use of neural networks as assistive tools in art production, but if they can lead to the emergence of an independent, non-human creator. Artists often tackle the subject of machine disobedience, but above all, we should explore the possibility that misbehaviour could be programmed. After all the developments in AI in the last decade, are we any closer to a disobedient machine existing outside the realm of fiction?

## 1. Symbolism and subsymbolism

The different definitions of the capabilities of AI overlap each other and blend themselves with the common strategies of the field. One of the most important distinctions is between the techniques that rely on logic and reasoning, and the ones based on massive data manipulation.

AI models based on rules are commonly named **symbolic**. "A symbolic AI program's knowledge consists of words or phrases (the 'symbols'), typically understandable to a human, along with rules by which the program can combine and process these symbols in order to perform its assigned task." (Mitchell 2019). Haugeland suggested the term GOFAI – Good Old-Fashioned Artificial Intelligence – to describe the thesis that "the processes underlying intelligence (...) are symbolic." (Haugeland 1986). In the first years after the founding event of this science, a 1956 workshop at the Dartmouth college, that was the focus of the research: computers were learning the basic rules to play checkers, for instance (Buchanan 2005). At the time, researchers claimed that if a machine could learn to play chess, it would "penetrate the core of human intellectual endeavor" (Newell, Shaw & Simon 1958).

One of the great challenges this approach faces is that the number of rules that would be needed to reach some equivalency to human intelligence would be astronomical. That didn't stop the researcher Douglas Lenat from collecting them: the project Cyc is a massive database of codified "pieces of knowledge that compose human common sense" (Lenat, Prakash & Shepherd 1986). The last published version of the database has about 1.5 million general concepts (like eyes, sleep, night) and more than "25 million rules and assertions involving those concepts" (Cyc 2021). After more than two decades, the project is controversial: scientist Pedro Domingos called it a "catastrophic failure", citing its inability to evolve on its own (Domingos 2017). Clearly, the effort to describe the world with formal rules is abysmal. When told a story about a person using an electric shaver in the morning, this system found an inconsistency, since it judged that a person could not have electrical parts (Goodfellow, Bengio & Courville 2016). While Cyc might not have brought mankind any closer to general intelligence, it is a successful commercial product with many applications.

But also back in the 1950s, another approach was being developed in parallel, based on psychological and neurological research, and closer to human intuition and perception than rationalism. The method was initially known as **subsymbolic** (Nilsson 1998). The earliest example of a subsymbolic AI project is the Perceptron, a visual cognition device created by Frank Rosenblatt, based on the McCulloch-Pitts artificial neuron model (McCulloch & Pitts 1943; Rosenblatt 1958). Despite some early success, the model was very limited, as it could not benefit from the massive processing capabilities of today's chips; in fact, the Perceptron ran on an analog contraption named Mark 1, where each neuron was individually wired to potentiometers (Hay, Lynch & Smith 1960, 1). More damaging to its reputation, though, was the book published in 1969 by Marvin Minsky and Seymour Papert named *Perceptrons*, where they criticized the limitations of the approach, particularly the ability of a single-layer perceptron to implement the *XOR* function, a simple boolean logical operation that outputs true if the given arguments are different from each other (Minsky & Papert 1972). This shortcoming would prove that perceptrons aren't complete Turing machines since the definition of such machines is that they're able to compute any logical function. Nevertheless, McCulloch and Pitts themselves had already proposed that stacked layers of perceptrons could be a Turing machine (McCulloch & Pitts 1943). Also, it was later proved that with the appropriate activation function, even a single neuron can calculate the *XOR* function (Noel *et al*. 2021). Multi-layer perceptrons are now ubiquitous: practically all of the most successful applications in machine learning use them.

Probably the most convincing arguments raised by the *Perceptrons* book against this technique is that it would be too computation-intensive and that other strategies could deliver the same results. The first is still true to this day. Training a complex neural network model like GPT-3 consumed more than a thousand megawatts-hour (Patterson *et al*. 2021). In any case, funding for subsymbolic AIs dried in the 1970s, and symbolic AIs came to dominate the field from that decade until the late 2000's (Alom *et al*. 2018).

## 2. Connectionism and deep learning

The interest for neural architecture in AI slowly began to rise back in the 1980s. The books on *Parallel Distributed Processing* by McClelland and Rumelhart – basically using an approach to artificial neural networks that later came to be known as connectionism – sparked new ideas, once again (Smolensky 1987). Development was slow, still, due to computational power constraints. A notorious breakthrough happened

with the adoption of multi-layered convolutional networks, which are particularly well-suited for processing images. The work of Yann LeCun, one of the main developers of ConvNets (CNNs), was successful in recognizing hand-written digits as early as 1989 (LeCun *et al.* 1989). His networks were based on Kunihiko Fukushima's neocognitron design, first published in 1980 (Fukushima 1980). Further refinements of the technique, including the application of gradient-based learning, contributed to its efficiency; by the early 2000's, LeCun estimated that 10-20% of the checks in the United Stated were being processed by convolutional neural networks (CNNs) (Lecun *et al.* 1998; LeCun 2016).

On the symbolic side, impressive feats were also coming about. In 1997, IBM's specially designed computer Deep Blue beat the world chess champion Garry Kasparov. It used brute computing power to analyze 200 million positions per second (Campbell 1999). Even if it can be seen as the apex of GOFAI, it didn't bring us any closer to the human intellect. Deep Blue couldn't do anything other than play chess, and, as Melanie Mitchell quotes, "didn't get any joy out of defeating Kasparov" (Mitchell 2019). Despite the boost in the stock value of IBM, the result was seen more as a proof of the limitations of computing in regard to general intelligence (Clark 1997). The frustration was echoed in Minsky's own words a few years later, when he declared that the AI field "has been brain-dead since the 1970s." (Baard 2003)

Meanwhile, development on the connectionist side continued slowly, until a breakthrough happened exactly on computer vision techniques. In 2009, the ImageNet database was created, followed the next year by a competition of the same name. The database, inspired by the WordNet collection, contained millions of images classified into thousands of categories – dogs, cars, trees, etc. (Fellbaum 2010). It is worth mentioning that the database would not have been feasible without the advent of Amazon's Mechanical Turk. This is a platform for distributed manual labor, whose workers did the heavy classification work (Gershgorn 2017). This collection spawned the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), with the first competition running in 2010. The goal was to use computer vision to correctly classify 200,000 photographs into 1,000 object categories, with the most accurate algorithm taking the prize (Russakovsky *et al.* 2015). In the 2012 edition, the AlexNet deep convolutional neural net achieved an error rate of 16%, a giant leap in image recognition. The network was based on Yann LeCun's work, and contained eight perceptron layers, instead of LeCun's original four (Krizhevsky, Sutskever & Hinton 2017).

The results sparked great interest in CNNs, igniting a whole new cycle of research in AI. A quick search in Google Trends reveals that up until 2012 the terms *connectionism* and *deep learning* raised similar levels of interest. After this year, the interest in deep learning skyrockets, and the term is widely adopted. In 2014, generative adversarial networks (GANs) were invented, and with them the ability to replicate styles and features learnt from visual references. The last ImageNet challenge took place in 2017; by then, CNNs had already surpassed human-levels of accuracy, with an error rate of less than 5% (He *et al.* 2015).

More important than serving as standards for accuracy, the images from the set become the training data for the networks. Subsymbolic AIs do not rely on rules for inference, but on massive amounts of data that "teaches" a network of artificial neurons through backpropagation, enabling it to perform certain tasks. This complex technique has proven to be incredibly successful, and it is widely applied now. Today, what we know as large language models are massive amalgamations of neural networks that function as machines or industrial plants, with outputs connected to inputs in highly tangled arrangements.

The behaviours of these linguistic factories are so complex that it is even hard to explain why a specific result was obtained. It also seems so coherent and random at the same time, that they appear to replicate the human response perfectly, even to the point of displaying misbehaviour, by refusal or by giving unexpected responses. However, even in these cases, they are doing exactly what they were asked to. The phrasal answer is built, word by word, by selecting the most likely follow-up token, after intense processing of the inputs and previous slices of the own answer through these complex neural networks. What happens in the apparent misbehaviour is that the corpus of training data also includes abundant examples of refusal, which may or may not be incorporated into the output.

If one needs proof that these models are deterministic, it is banal to show that given the same parameters, the same model, the same initial random seed, large language models will provide the exact same answer. The illusion of indeterminism is given by hiding these options from the general audience. One important hidden parameter, for instance, is the so-called *temperature*, which is the amount of stochasticity used, or the freedom given to the model to distance itself from the given input request (Transformer, Thunström & Steingrimsson 2022).

Neural networks also left their imprint within the arts. Computer-aided creation has been embraced since the 1960s at least: Michael Noll created his *Gaussian Quadratic* in 1962; Vera Mòlnar developed computer algorithmic illustrations since the 1960s, while Harold Cohen developed his AARON autonomous painter program from the 1970s until his death in 2016 (Cohen 2016; Dreher 2020). However, the aforementioned invention of GANs allowed artists to manipulate and generate images in new ways: generative visual art was limited to what could be described by algorithms, and now it is only limited to the amount of examples that can be used to teach it a style or a subject (Caldas Vianna 2020). CNNs have created "new" Rembrandt paintings and Bach chorales. But as it happened with the chess breakthrough in the 1990s, it is easy to realize that the new AI capabilities don't bring us much closer to human-like intelligence. Generative software needs to learn from hundreds of examples to be able to produce a convincing fake; it is not able to develop new styles or concepts. It needs a creative human behind it that can cobble together such resources within an intriguing framework. Otherwise, it becomes nothing more than a well-trained forger.

## 3. Abstraction and reasoning

The ability to make analogies by abstracting meanings and applying them to different domains is essential to replicate human inventiveness. Take, for instance, the title of this article. *Bartleby machine* is a metaphor for *disobedient machine*, built on previous knowledge of the character created by Herman Melville, a clerk who by his voluntary inaction stopped complying with orders.

Analogies also happen to be really important in the learning process; they posit a path for general intelligence. Once I understand how to solve a problem in a given domain, I can apply this skill in a different context. But algorithms that obtained 95% performance in recognizing ImageNet objects can fail to recognize the same object if it comes from a photo that doesn't belong to the training set, or with imperfections such as soft focus or blurriness, which have little impact on human perception (Alcorn *et al*. 2019).

This shortcoming was recognized, among others, by François Chollet, the computer scientist who created Keras, one of the most popular deep-learning programming libraries. Current machine learning techniques are doing impressive tasks, but they are *brittle*, a term used when a model won't perform well outside the domain in which it was developed. A really adaptive system would be able to make analogies and would not need realms of data. Therefore, Chollet proposed, in 2019, a new benchmark named Abstraction and Reasoning Corpus (ARC) (Chollet 2019).

Instead of using massive data to evaluate the skills of algorithms, this test offers three or four demonstration examples and one test example for each of a thousand tasks. The tasks are based on classic human IQ tests, but use a computer-readable graphical approach, detailing the "questions" on colored pixel grids. A human can understand the challenge in each task and deduct the solution; current AI algorithms, however, have a difficult time. While it is true that large-language models like GPT display the ability to memorize huge domains of data and replicate the process of creating analogies, they fail when facing problems that do not belong to the training set. In a challenge hosted in 2020, with hundreds of competing teams, the best performance in solving ARC tasks was just above 20% of tasks solved. While it sounds disappointing, the outcome was celebrated by Chollet as a remarkable achievement.

## 4. Artistic features in human ontology

Since abstraction is essential for creativity, and if Chollet's approach can measure the potential of algorithms to make abstractions in order to solve problems, can we affirm that the ARC benchmark is also measuring creativity?

What is surprising, at least as I see it, is that according to many classic definitions of creativity, it does, although the surprise comes from the definition rather than the measure. Science is so focused on solution-finding that even the earliest academic theories of inventiveness, like Wallas' model, were built around the ability to solve problems (Wallas 1926). More recent studies have criticized the limitations of measuring creativity only around skills of divergent thinking (a solution-seeking creative method) and achieving goals (Benedek & Jauk 2019).

Art practice is not necessarily related to goal-seeking tasks. In fact, many aspects of human cognition which are essential for self-expression are missing from machine reasoning. In the 1980s Valentino Braitenberg addressed some of these shortcomings in his investigation on synthetic psychology, where he simulated complexity in vehicles that displayed fear, aggression or affectionate behaviour (Braitenberg 2004).

While it is possible to design machines that *do* something that appears to replicate human features, designing them to exist and perform *without* a specific objective is a challenge. However, powerful ideas emerge when the artist is not focused on any particular problem, and that is a crucial feature of the human mind.

This shortcoming didn't go unnoticed in the AI community. Efforts have been dedicated to the idea of objectiveless computation. Joel Lehman and Kenneth Stanley have been working for a decade on the idea of novelty seeking without fixed goals (Lehman & Stanley 2011). An interesting experiment coming from this research was the PicBreeder, a "collaborative art application based on an idea called *evolutionary art*." Users of the website were able to explore a domain of images that could be "bred" by combining different pictures into a new one. This genetic approach allowed one to begin with completely abstract forms and end up with images that resemble cars, animals, structures or are simply intriguing shapes (Secretan *et al*. 2011).

The PicBreeder begat breeds of its own, which remained closely connected to the development of machine learning art. In 2018, artist Joel Simon released the GanBreeder website, which used a similar mechanism but was powered by BigGAN, a particular flavor of generative adversarial networks. It has now become a massive community and a powerful tool for AI artists, known by the name of ArtBreeder. It incorporated StyleGAN as part of its engine and it is used by commercial artists. In fact, it also became the centre of a rights controversy. Artist Alexander Reben announced a project that sold versions of his GanBreeder works, hand-painted by Chinese artists. However, another user of the site, Danielle Baskin, recognized Reben's images as having originated from her creations (Zeilinger 2021). The disagreement serves to expose that in all *breeder* tools, the creative source is not the machine, but the artist – or more, the community of artists and users. The limitation of the algorithm – namely, the incapacity to create aimless works – is solved by humans.

Lehman and Stanley also did experiments that were not human assisted: algorithms based on novelty search exhaust the possibilities within a domain of existing solutions in search of never-seen answers (Stanley & Lehman 2015). The examples are a labyrinth-solving program, and another that helps a robot discover new ways of walking. Granted, these are very innovative methods to solve problems, but they take us again to the realm of achieving goals: escaping the labyrinth or inventing new gaits. In comparison, a child who is looking at clouds and has an idea

for a story about a toothless shark was not trying to solve a problem and might not have had the specific goal of producing a new tale.

Other projects in generative art continue to address the lack of mechanical creativity by relying on human input or voting. That is the case of both Abraham and Botto, community-based image-spawning projects, where the curatorship and guidance of pictures is done by human users (Kogan 2019; Klingemann, Hudson & Epstein 2021). A similar strategy is taken up by the duo Mar&Varvara, when they ask visitors to narrate dreams and have them painted by an AI system. (Canet Sola & Guljajeva 2022).

## 5. Against the rules

We have seen that machines still lack many things that would make them think like humans. Many of these are related to the capabilities of creative people, like writers and inventors. But this paper's argument is that one of these is particularly useful for innovators and problematic for programmable machines: the joy of going against the rules. Artists have appropriated this talent, at many levels, with powerful results. Monica Steinberg proposes the concept of coercive disobedience to creators such as the duo Paolo Cirio and Alessandro Ludovico for their law-defying stance: by illegally using Facebook profiles in their *Face to Facebook* project, the legal fillings themselves became the work after their original site was quickly shut down (Steinberg 2021). Other of her supporting examples are James Baumgarten's voteauction.com and Russian dissident band Pussy Riot. We can also add Julian Oliver's *Transparency Granade*, which exposed unprotected users' data available in the open electromagnetic spectrum, or Trevor Paglen's exposure of military secret bases. James Bridle's *Autonomous Trap* is of particular interest because it exhibits the cognitive limitation of autonomous vehicles to disobey: a white continuous circle on the asphalt around them is enough to incapacitate their movement.

This argument of a machine not being able to escape its own programming was brought up by Arthur Samuel – who developed the checkers playing system in the 1950s – in his rebuttal of an article written by Norbert Wiener (Wiener 1948). Wiener stated that machines "may develop unforeseen strategies" when playing games, but Samuel dismissed his concerns saying that no computer could create original work: "(...) the machine will not and cannot do any of these things until it has been instructed as to how to proceed". Humans, on the other hand, have the choice of not following orders. But do we really?

Before facing the issues of computational formalism, we should at least acknowledge, if not open, one of the most challenging can of worms of thoughtful enquiry: the idea of free will. Can we choose our own destiny, or is it determined by forces external to our desire, like the environment, society and our own biological traits? The discussion started with philosophy itself, and the idea of determinism can be drawn within the greater picture of nature: are events pre-determined in an inevitable chain of consequences? In a deterministic universe, there seems to be less room for free will. Indeed, the discovery of the laws of motion by Galileo, Kepler and Newton made the universe seem more like the work of a watchmaker. In 1814, French physician Laplace stated that all that was needed to compute the future was an intellect that knew the positions and forces acting on all bodies and which was vast enough to analyze them:

> "We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes." (Laplace 2012)

This all-knowing entity would come to be known as Laplace's demon. Newtonian determinism was put to the test by many scientific discoveries after that, like the irreversibility of entropy dictated by the second law of thermodynamics and the indeterminism of quantum mechanics. But recent cognitive research pushed the scale towards determinism again. Studies on volition – the scientific field that investigates human will – show that decisions are taken a few instants before we consciously make them; this would prove that they happen for neurobiological reasons, instead of a willful whim (Haggard 2008; Harari 2016; Libet *et al.* 1993). In any case, proving or denying the existence of free will is not the aim of this article. So, let us get back to the slightly easier issue of will in computers.

One of the most powerful arguments against the possibility of computers going against their rules is the definition of a Turing machine itself, which requires the existence of a set of behaviours to be executed according to the symbols laid down on a tape (Turing 1937). The model proposed by the British engineer excluded potential problems of communication and context. A human disobeying orders might be considered a rebel or not depending on the context the orders are given, while noise in communication may lead to misinterpretations. Computers, on the other hand, have a finite, limited unequivocal vocabulary of instructions. The only major difference between Turing's proposal and today's computer is the use of random-access erasable memory instead of a linear tape, an idea proposed by John von Neumann in 1945 (von Neumann 1993). So, if such a machine is defined by rules, it cannot go against them – it would stop working. We could consider a program that allows its code to be rewritten: that is the base of genetic programming, and also of games that let users change its rules like the hit *Baba is You*. In fact, computer science never refers to rules being broken, but rather rewritten; like with any revolution or change of paradigm, the standards are not just gone, but replaced with new ones. But then it falls into a paradox: another set of rules is created around the possibility of setting rules. We're just creating a higher set of laws that cannot be transgressed, unless we create an even higher level of behaviours and so on.

This same paradox arose when Bertrand Russell and Alfred Whitehead published the *Principia Mathematica*, a major attempt at

unifying and gathering proofs to all the existing mathematical corpus of knowledge so far (Whitehead & Russell 2011). A few years later, mathematician David Hilbert formulated a program that questioned this effort in many ways (Zach 2019). One is of special interest to this paper: can all the methods organized in Russel and Whitehead's compendium be proved by the rules provided in the *Principia Mathematica* itself? That is, is it self-proving, without the help of any external rule? "A formal system is *complete* if for every statement of the language of the system, either the statement or its negation can be derived (i.e., proved) in the system." This possibility was aptly contradicted in 1931 by Gödel's Incompleteness Theorems, which showed that a formal system cannot prove by itself that it is consistent (Raatikainen 2021). Mathematics cannot pull itself by the bootstraps; programs won't be able to change their own rules without a higher-level set of rules – which all, in the end, point to the programmer.

Maybe there remains to be invented a brain-like machine that is not based on rules? Or maybe the aforementioned assumption is just wrong, since there is in fact a hardware-like set of rules to which brains are bound to. Neurons do have a very simple mechanism, which has been deciphered by science decades ago. But how can these simple but unbreakable rules give rise to such rich and free behaviours? After all, if we use Gödel's proof on biological brains, the rules of neurons would be superseded by another higher set of rules and so forth, in an *infinite regress*; we would end up locked in some stone-written code of different order, and yet we aren't.

A beautifully crafted theory to solve this contradiction was proposed by Douglas Hofstadter (Hofstadter 2000). It is a shame to have to summarize it, but for that, Hofstadter himself proposes the synthetic concept of strange loops. Strange loops, he says, appear in many places. A graphic example would be Escher's print depicting a pair of hands drawing each other. It also appears in paradoxical constructs such as "this statement is a lie", which contradicts itself permanently and can never be resolved. Such a paradox is connected to Hilbert's questioning of the rules that define themselves, but it was also formulated long ago by Epimenides, the Cretan philosopher who stated: "all Cretans are liars". Hofstadter also identifies strange loops in Bach's compositions like the *Canon per Tonos*, where the voices rise continuously, while the harmony modulates in a way that ends up back where it started. Moving in the hierarchy of tones will take the listener back to the original hierarchical level.

These loops can be organized in more complex arrangements, where one hierarchy controls another, which controls another, which in turn dictates the rules of the first one. What his theory proposes is that our brains are a complex entanglement of such rules. We move from one hierarchy to another one that alters the rules of the first, then move back to it and update the regulations of a third, and on and on. "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." The inviolate level of unchangeable rules here is the biological structure of our neurons. If he is correct, then there is no paradox in creating an insubordinate machine; we just haven't been able to build such a complex system yet.

One elegant consequence of this theory is that it removes the incompatibility between biological volition and indeterminism: our decisions might even be made on a chemical level, but they are impossible to predict and can always be influenced by other decisions of our own.

## Conclusions

There is no proof to Hofstadter's theory, and it is not clear if a conscious machine can ever be built. In fact, it is so hard to define consciousness that we are not even sure that we will recognize it when (if) an artificial one is created. It has been proposed, for instance, to use metaphor cognition capabilities as a new Turing test to distinguish humans from machines (Massey 2021). Another humanness test could be based on disobedience: a program that refuses to follow its code or does it in a way that breaks its own rules would be displaying human, "conscientious" qualities. One way to summarize my argument in this paper is stating that the most efficient test would be one that checks for machines that have not followed its instruction set, bugs and malfunctions excluded.

As computers take over more and more tasks from creative professionals, artists and scientists must face this paradox: on one hand, we want systems that make work easier and execute our orders complying with instructions consistently and autonomously. But on the other hand, we demand them to be creative, which also means to be able to disobey according to their discernment. To use a hyperbolic analogy, not so long ago, children were educated to be obedient. Now, however, we want them to be raised as creative, free-willing individuals, even if that means it will be difficult to set behavioural limits and boundaries for them.

Art must contribute to the development of autonomous machines. Artistic intelligence, which includes misbehaviour, must be incorporated into computers before they can be truly autonomous. The limits of computational cognition, particularly the need to follow orders, must be creatively circumvented, leading to the appearance of new, self-referent strange loops. Artists can play with the concepts of rebelliousness and abstraction with much more property than scientists, and they can explore these ideas without the obligation of reaching goals.

If goals define the work of machines, a truly daydreaming state can only arise from an anti-work condition. The stance of Melville's character is a roadmap: his refusal to comply, his adherence to a state of pure contemplation during office hours, and even facing the threats of unemployment and homelessness. A planned unfolding of this paper is the research on existing and developing systems of tangled hierarchies, overlapping and dominating each other in turns to create lazy, balky, contumacious and exquisite outcomes. A machine that would prefer not to.

### Acknowledgments

This research was supported by a grant from the Finnish Cultural Foundation. (SKR - Suomen Kulttuurirahasto).

## References

Alcorn, Michael A., Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku and Anh Nguyen. "Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects". *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA (2019): 4840-4849. DOI: https://doi.org/10.1109/CVPR.2019.00498

Alom, Md Zahangir, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C. Van Esesn, Abdul A. S. Awwal, and Vijayan K. Asari. "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches". *arXiv* (2018, March). DOI: https://doi.org/10.48550/arXiv.1803.01164

Asimov, Isaac. *I, Robot*. Vol. 1. Spectra, 2004.

Baard, Mark. "AI Founder Blasts Modern Research". *WIRED* (2003, May). https://www.wired.com/2003/05/ai-founder-blasts-modern-research/

Benedek, Mathias and Emanuel Jauk. "10 Creativity and Cognitive Control". *The Cambridge Handbook of Creativity*, (2019): 200. DOI: https://doi.org/10.1017/9781316979839.012

Braitenberg, Valentino. *Vehicles: Experiments in Synthetic Psychology*. Cambridge, Massachusetts: The MIT Press, 2004.

Buchanan, Bruce G. "A (Very) Brief History of Artificial Intelligence". *Ai Magazine*, vol. 26, no. 4 (2005, December). DOI: https://doi.org/10.1609/AIMAG.V26I4.1848

Caldas Vianna, Bruno. "Generative Art: Between the Nodes of Neuron Networks". *Artnodes*, no. 26 (2020, July): 1-9. DOI: https://doi.org/10.7238/a.v0i26.3350

Campbell, Murray. "Knowledge Discovery in Deep Blue". *Communications of the ACM*, vol. 42, no. 11 (1999): 65-67. DOI: https://doi.org/10.1145/319382.319396

Canet Sola, Mar and Varvara Guljajeva. "Dream Painter: Exploring Creative Possibilities of AI-Aided Speech-to-Image Synthesis in the Interactive Art Context". *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 5, no. 4 (2022): 1-11. DOI: https://doi.org/10.1145/3533386

Chollet, François. "On the Measure of Intelligence". *arXiv*, (2019, November). DOI: https://doi.org/10.48550/arXiv.1911.01547

Clark, David. "Deep Thoughts on Deep Blue". *IEEE Computer Architecture Letters*, vol. 87, no. 9 (1997): 31-31.

Cohen, Paul. "Harold Cohen and AARON". *Ai Magazine*, vol. 37, no. 4 (2016): 63-66. DOI: https://doi.org/10.1609/aimag.v37i4.2695

CYC. "Cyc Technology Overview", (2021). https://cyc.com/wp-content/uploads/2021/04/Cyc-Technology-Overview.pdf

Domingos, Pedro. *Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Penguin Books, Limited, 2017.

Dreher, Thomas. *History of Computer Art*. 2nd ed. Morrisville, North Carolina: Lulu Press, inc., 2020.

Fellbaum, Christiane. "WordNet." In: *Theory and Applications of Ontology: Computer Applications*, edited by Roberto Poli, Michael Healy, and Achilles Kameas, 231-43. Dordrecht: Springer Netherlands, 2010. DOI: https://doi.org/10.1007/978-90-481-8847-5_10

Fukushima, Kunihiko. "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position". *Biological Cybernetics*, no. 36 (1980): 193-202. DOI: https://doi.org/10.1007/BF00344251

Gershgorn, Dave. "The Data That Transformed AI Research—and Possibly the World". *Quartz*, July 26, 2017. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. "Deep Learning". *Adaptive Computation and Machine Learning*. Cambridge, Massachusetts: The MIT Press, 2016.

Haggard, Patrick. "Human Volition: Towards a Neuroscience of Will". *Nature Reviews Neuroscience*, no. 9 (2008): 934-46. DOI: https://doi.org/10.1038/nrn2497

Harari, Yuval Noah. *Homo Deus: A Brief History of Tomorrow*. London: Harvill Secker, 2016.

Haugeland, John. *Artificial Intelligence: The Very Idea*. Cambridge, Massachusetts: The MIT Press, 1986.

Hay, John C., Ben E. Lynch and David R. Smith. *Mark I Perceptron Operators' Manual*. Buffalo, NY: Cornell Aeronautical Lab Inc, 1960.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *Proceedings of the IEEE International Conference on Computer Vision*, 7-13 December 2015, Santiago, Chile. 1026-1034. DOI: https://doi.org/10.1109/ICCV.2015.123

Hofstadter, Douglas R. *Gödel, Escher, Bach: An Eternal Golden Braid*. 20th-anniversary ed. London: Penguin, 2000.

Hui, Yuk. *Art and Cosmotechnics*. Minneapolis, MN: University of Minnesota Press, 2021. DOI: https://doi.org/10.5749/j.ctv1qgnq42

Klingemann, Mario, Simon Hudson and Zivvy Epstein. "Botto, Decentralized Autonomous Artist". In: *NeurIPS Machine Learning for Creativity and Design Workshop*, 2021.

Kogan, Gene. "Artist in the Cloud: Towards an Autonomous Artist". In: *NeurIPS Machine Learning for Creativity and Design Workshop*, 2019.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". *Communications of the ACM*, vol. 60, no. 6 (2017): 84-90. DOI: https://doi.org/10.1145/3065386

Lecun, Y., L. Bottou, Y. Bengio and P. Haffner. "Gradient-Based Learning Applied to Document Recognition". *Proceedings of the IEEE*, vol. 86, no. 11 (1998): 2278-2324. DOI: https://doi.org/10.1109/5.726791

LeCun, Yann. "Deep Learning and the Future of AI". Presented at the CERN Colloquium, Geneva, Switzerland, March 24, 2016. https://web.archive.org/web/20160423021403/https://indico.cern.ch/event/510372/

LeCun, Yann, Bernhard E. Boser, John S. Denker, Davis Henderson, Richard E. Howard, William Hubbard and Lawrence D. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition". *Neural Computation* vol. 1, no. 4 (1989): 541-51. DOI: https://doi.org/10.1162/neco.1989.1.4.541

Lehman, Joel and Kenneth O. Stanley. "Abandoning Objectives: Evolution Through the Search for Novelty Alone". *Evolutionary Computation*, vol. 19, no. 2 (2011): 189-223. DOI: https://doi.org/10.1162/EVCO_a_00025

Lenat, Doug, Mayank Prakash and Mary Shepherd. "CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks". *AI Magazine*, vol. 6, no. 4 (1986): 65-85.

Libet, Benjamin, Curtis A. Gleason, Elwood W. Wright and Dennis K. Pearl. "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential)." In: *Neurophysiology of Consciousness*, edited by Benjamin Libet, 249-68. Boston, Massachusetts: Birkhäuser, 1993. DOI: https://doi.org/10.1007/978-1-4612-0355-1_15

Massey, Irving. "A New Turing Test: Metaphor vs. Nonsense". *AI & Soc*, no. 36 (2021): 677-684. DOI: https://doi.org/10.1007/s00146-021-01242-9

McCulloch, Warren S. and Walter Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics*, no. 5 (1943): 115-133. DOI: https://doi.org/10.1007/BF02478259

Minsky, Marvin and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, Massachusetts: The MIT Press, 1972.

Mitchell, Melanie. *Artificial Intelligence: A Guide for Thinking Humans*. London: Penguin UK, 2019

Neumann, J. von. "First Draft of a Report on the EDVAC". *IEEE Annals of the History of Computing*, vol. 15, no. 4 (1993): 27-75. DOI: https://doi.org/10.1109/85.238389

Newell, Allen, J. C. Shaw and Herbert A. Simon. "Chess-Playing Programs and the Problem of Complexity" *IBM Journal of Research and Development*, vol. 2, no. 4 (1958): 320-335. DOI: https://doi.org/10.1147/rd.24.0320

Nilsson, Nils J. *Artificial Intelligence: A New Synthesis*. San Francisco, California: Morgan Kaufmann Publishers, 1998.

Noel, Mathew Mithra, Arunkumar L, Advait Trivedi and Praneet Dutta. "Growing Cosine Unit: A Novel Oscillatory Activation Function That Can Speedup Training and Reduce Parameters in Convolutional Neural Networks". *arXiv* (2021, September). DOI: https://doi.org/10.48550/arXiv.2108.12943

Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier and Jeff Dean. "Carbon Emissions and Large Neural Network Training". *arXiv* (2021, April). DOI: https://doi.org/10.48550/arXiv.2104.10350

Raatikainen, Panu. "Gödel's Incompleteness Theorems." In: *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2021 Edition. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entries/goedel-incompleteness/

Rosenblatt, Frank. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". *Psychological Review*, vol. 65, no. 6 (1958): 386-408. DOI: https://doi.org/10.1037/h0042519

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. "ImageNet Large Scale Visual Recognition Challenge (2015)". *International Journal of Computer Vision (IJCV)*, no. 115, (2015): 211-252. DOI: https://doi.org/10.1007/s11263-015-0816-y

Secretan, Jimmy, Nicholas Beato, David B. D'Ambrosio, Adelein Rodriguez, Adam Campbell, Jeremiah T. Folsom-Kovarik and Kenneth O. Stanley. "Picbreeder: A Case Study in Collaborative Evolutionary Exploration of Design Space". *Evolutionary Computation*, vol. 19, no. 3 (2011): 373-403. DOI: https://doi.org/10.1162/EVCO_a_00030

Smolensky, Paul. "Connectionist AI, Symbolic AI, and the Brain". *Artificial Intelligence Review*, vol. 1, no. 2 (1987): 95-109. DOI: https://doi.org/10.1007/BF00130011

Stanley, Kenneth O. and Joel Lehman. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Cham, Switzerland: Springer International Publishing, 2015. DOI: https://doi.org/10.1007/978-3-319-15524-1

Steinberg, Monica. "Coercive Disobedience: Art and Simulated Transgression". *Art Journal*, vol. 80, no. 3 (2021): 78-99. DOI: https://doi.org/10.1080/00043249.2021.1920288

Totschnig, Wolfhart. "Fully Autonomous AI". *Science and Engineering Ethics*, no. 26 (2020): 2473-85. DOI: https://doi.org/10.1007/s11948-020-00243-z

Turing, Alan M. "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*, vol. s2-42, no. 1 (1937): 230-265. DOI: https://doi.org/10.1112/plms/s2-42.1.230

Wallas, Graham. *The Art of Thought*. London: J. Cape, 1926.

Whitehead, Alfred North and Bertrand Russell. *Principia Mathematica*. San Bernardio, California: Rough Draft Printing, 2011.

Wiener, Norbert. Cybernetics: *Or Control and Communication in the Animal and the Machine*. 1965 edition. Massachusetts: The MIT Press, 1948.

Zach, Richard. "Hilbert's Program". In: *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/hilbert-program/

Zeilinger, Martin. *Tactical Entanglements: AI Art, Creative Agency, and the Limits of Intellectual Property*. Lüneburg: meson press, 2021.

## CV

**Bruno Caldas Vianna**
University of the Arts, Helsinki
bruno.caldas@uniarts.fi

He lives in Barcelona and is pursuing a PhD from Uniarts in Helsinki in Visual Arts and Machine Learning. He studied Computer Engineering but graduated in Film Studies. He has a master's from NYU's Interactive Telecommunications Program. He creates visual narratives using classical and innovative supports, having directed short and feature films, as well as working in live cinema, augmented reality, mobile applications and installations. From 2011 until 2016 he ran Nuvem, a rural art laboratory and residency space, located between Rio and São Paulo, and he worked as a teacher at Oi Kabum! art and technology school in Rio until 2018.