Artificial Intelligence and Creativity: an Aesthetic Examination of Computer-generated Art

Raos, Robert

University of Rijeka

Faculty of Humanities and Social Sciences

Department of Philosophy

Robert Raos

Artificial Intelligence and Creativity: An Aesthetic Examination of Computer-Generated Art

Rijeka, 2020.

University of Rijeka

Faculty of Humanities and Social Sciences

Department of Philosophy

Student Name: Robert Raos
(JMBAG: 0009080575)

Artificial Intelligence and Creativity: An Aesthetic Examination of Computer-Generated Art

Bachelor's degree programme: Philosophy

Mentor: doc. dr. sc. Iris Vidmar Jovanović

Rijeka, 2020.

Abstract

The idea that artificial intelligence could produce art is met with numerous philosophical challenges. Among them, one of the chief questions is how an agent with no internal mental states could be an author. Given that most theories of art presuppose authorial intent, this paper attempts to reconcile it with computer-generated art. Namely, my main thesis is that hypothetical intentionalism enables the notion of AI being an author. To reasonably defend the theory of hypothetical intentionalism, I first need to eliminate the viability of actual intentionalism, as well as argue that other alternatives -- such as death of the author and anti-intentionalism -- are not as attractive.

My argumentation will be divided into two chapters, although the points presented in each are mutually complementary. The first chapter will tackle the discourse of authorial intent. More specifically, it will critique various accounts of intentionalism. The ultimate claim of the chapter will be that hypothetical intentionalism is both the most philosophically sound position and the most supportive of computer-generated art. The second part will review the current technological state of artificial intelligence and contextualize it in a philosophically meaningful way. Most importantly, it will demonstrate that -- presupposing the truth of hypothetical intentionalism -- computer authors are able to meet all the criteria necessary for art generation.

This paper presents authorial intent as an interactive, malleable entity rather than a one-way imposition. Consequently, it seeks to trigger contemplation of how we approach technology and artwork. Meaning is characterized as a public object that can be mass-produced by non-human subjects.

Keywords: artificial intelligence, authorial intent, intentionalism, author, meaning, machine learning, neural network

Table of Contents

1.) Authorial Intent

1.1 Introduction

The first half of my essay will explore the concept of authorial intent. Namely, I will position myself in the discussion surrounding intentionalism by repudiating theories of actual intentionalism and defending hypothetical intentionalism. The resulting postulation of a non-existent author who is imagined by art interpreters will enable me to qualify non-thinking entities (computers) as authors.

1.2 The Role of the Author

In his book *Philosophy of Art*, Noël Carroll takes an analytic approach to Art Theory. He ventures to discover the conditions that are both necessary and sufficient to qualify an entity as an artwork. Rather than favoring a single account, he provides an overview of the most prominent theories of art. Among the various definitions he puts forth, a common thread emerges: *authorial intent*. The intention to elicit a specific response in an audience seems to be a critical component of most theories of art.

Take, for instance, the neo-representational theory. It posits that, in order for a work to become an artwork, it needs to possess the property of "aboutness" or, in other words, it needs to have "semantic content" (Carroll 27). Crucially, this representative relation must be intentional:

> A visual design x pictorially represents y (an object, person, place, action, event or another visual design) if and only if (1) has the **intended** capacity to cause a normal percipient to recognize y in x. (Carroll 45)[1]

Similarly, as the name might suggest, the expressivist theory of art requires the artist to *express* their individualized feelings via art. Expressivism postulates that what makes art a

[1] This paper uses MLA in-text citations

unique practice is its focus on the emotional intensity involved in both the creation and reception of artworks. Artists are expected to intentionally transport their feelings through an artistic medium:

> "x is a work of art if and only if x is (1) an **intended** (2) **transmission** to an audience (3) of the self-same (type identical) (4) individualized (5) feeling state (emotion)(6) that the artist experienced (himself/herself) (7) and clarified (8) by means of lines, shapes, colors, sounds, actions and/or words." (Carroll 65)

Furthermore, theories of art that center the form of a work are also predicated upon the concept of authorial intent. To disqualify objects that possess significant form but are not art, the notion of intelligent design must be introduced to rescue formalism from being too inclusive:

> "x is a work of art if and only if x is **designed primarily** in order to possess and to exhibit significant form." (Carroll 115)

Even the aesthetic theory of art, which seems to solely require an object to generate aesthetic appreciation in an audience to be an artwork, cannot escape the concept of intentionality. Since many things have the capacity to generate aesthetic experiences in people (be they positive or negative), ranging from vomit on a city sidewalk to a Coca-cola advertisement, sentient decision-making helps delineate what does and does not constitute art:

> "x is an artwork if and only if (1) x is produced with the intention that it possess a certain capacity, namely (2) the capacity of affording aesthetic experience." (Carroll 162)

As Noël Carroll takes us on a trip through the history of aesthetics, it becomes apparent that a majority of the relevant theories of art demand authorial intent. Be it semantic content,

expressive content, contemplation of a significant form, or an aesthetic experience, a specific mental state needs to travel from the mind of the author to the mind of the recipient in order for a work to be promoted to the status of artwork.

How, then, does one even entertain the possibility of computer-generated art? Currently, we lack the technological sophistication to engineer programs with any internal mental states, let alone the desires, hopes, and dreams typically associated with artists. The following sections will probe deeper into what it means to have an internal mental state, how meaning is embedded in symbolism, and will question whether the idea of authorial intent is even tenable.

1.3 Actual Intentionalism

1.3.1 Intentionalisms

What actual intentionalist theories of art have in common is their shared assertion that the author has some degree of authority in determining the meaning of an artwork. Consulting him may illuminate the work's correct interpretation. The following section will examine the most direct variant of intentionalism: actual intentionalism. In particular, I will challenge its conclusions by questioning its epistemic viability.

1.3.2 Extreme Intentionalism

In *Validity in Interpretation,* E.D Hirsch championed his brand of intentionalism, which would later be characterized as "extreme intentionalism" by his adversaries[2]. Written in 1967, it responded to the attempts of authors such as T.S Eliot to separate the author's intentions from the work they spawned. One of his strongest arguments for the necessity of the author came in the form of a rhetorical question:

A word sequence means nothing in particular until somebody either means something by

[2] Swirski (8)

> it or understands something from it. There is no magic la of meanings outside of human consciousness. Whichever meaning is connected to words, a person is making the connection, and the particular meanings one lends to them are never the only legitimate ones under the norms and convictions of his language. One proof that the conventions of language can sponsor different meanings from the same sequence of words can and do disagree. When these disagreements occur, **how are they resolved**? (4)

Since different interpreters sometimes provide conflicting accounts of a work's meaning, Hirsch claims that the only agent capable of resolving the dispute is the author. "Whenever meaning is attached to a sequence of words it is impossible to escape an author." (Hirsch 5) According to this radical form of intentionalism, the only correct interpretation is the one that corresponds to the intended meaning of the author. However, this approach to intentionality faces a serious epistemic obstacle: how does one know the original intent of the author? This problem was perhaps first articulated by K.Wimsatt and M. C. Beardsley in their essay *The Intentional Fallacy*:

> One must ask how a critic expects to get an answer to the question about intention. How is he to find out what the poet tried to do? If the poet succeeded in doing it, then the poem itself shows what he was trying to do. And if the poet did not succeed,
> then the poem is not adequate evidence, and the critic must go outside the poem for evidence of an intention that did not become effective in the poem. (469)

This disjunction enables us to eliminate the need for an author's external interpretative input. Should her art be well-designed, the intended meaning may be extracted from her work, making the author's post-hoc input redundant. If, on the other hand, audiences fail to recognize the intended meaning of an artwork, the author was ineffective at communicating it in the first

place, making the author's intention insufficient for its recognition. Thus, external evidence of intended meaning (such as personal letters or tweets) is either unnecessary or insufficient. Essentially, the only way to gauge 'what the poet tried to do" is to read the poem. The intention of the author cannot be accessed directly (468). To bolster the attack presented by the aforementioned authors, I will apply the concept of the Black Box to this discussion.

In 1941, the mathematician Wilhelm Cauer introduced the concept of a "Black Box", which was later appropriated by philosophy. It refers to a hypothetical system whose inputs and outputs are observable, but internal mechanisms are not (43). While this model was traditionally used in computer science to develop reverse-engineering techniques, it can be useful in other fields. For instance, political scientists may borrow the concept to discuss governments or organizations whose inner workings are unknown. A philosopher may use a "black box" model to describe how an individual perceives others. While one is certain that she experiences internal mental states, she cannot know that others do as well. After all, everything we observe in other people amounts to inputs and outputs. This skepticism about the subjective experiences of other people is explored most thoroughly by philosophers of mind. The polemic regarding the so-called "hard problem of consciousness" is elegantly summarized by Weisberg (ch. 1-2). In short, it is unclear why the neural circuitry of humans generates internal mental states. The biological mechanisms that engender our behavior could, theoretically, occur without a phenomenological interface. It is not unreasonable, then, to be skeptical of others' alleged internal realities.

Even if, however, one were to generously sidestep this philosophical conundrum, the mental states of the authors would still not be accessible, which calls into question the alleged power they hold over the art they spawned. Outside of fringe philosophy departments, the world has embraced empiricism and the scientific method. The properties of a material object are discovered by observing it. Artworks are material objects. Much to the dismay of some contemporary philosophers of art, there is no metaphysical umbilical cord connecting the artist

to his creation. The mental states of the artist are not empirically observable. Theoretically, one might record the electrical impulses of an author's brain during the creative process, but his subjective "qualia" would still remain a black box. Conversely, once an artwork is created, it becomes a physical entity subject to examination. Readers may attempt to ascertain what the mental states of the author were at the time of creation, but they can never test whether their hypotheses correspond to his actual mental state since it is, as previously argued, opaque. Moreover, most authors of the art we appreciate are dead. They do not exist. Their mental states do not exist. One cannot test and measure something that does not exist. Therefore, direct observation of an author's qualia is impossible. Since the author's mental states cannot be objectively excavated, extreme intentionalism is untenable.

1.3.3 Moderate Intentionalism

Some intentionalists attempted to make the theory more palatable by reducing the strength of its claims. There are multiple ways in which its proponents attempted to do this, but for the purposes of this essay, I shall tackle Peter Swirski's version of moderate intentionalism:

> "A moderate intentionalist's contention—albeit one fraught with repercussions for critics inclined to level all interpretations in the name of textual jouissance—is that authorial intentions must inform (rather than determine) any interpretation of a literary work or an artwork." (141)

One of the problems with this amended definition is that it does not address the primary weaknesses of the theory. Namely, the epistemically dubious access interpreters have to authorial intentions remains the same regardless of whether they determine or inform an interpretation of a literary work. Consequently, Swirski fails to differentiate moderate intentionalism from extreme intentionalism (in this regard) by adjusting the definition. He does,

however, attempt to tackle the epistemic and ontological challenges to his theory in the following quote:

> "the potential empirical difficulties of elucidating the fact of the matter—the executive intentions of the creator—do not negate that there is a fact of the matter" (80)

I offer two responses to this argument. Firstly, empirical inquiry into the intentions of the creator is not only difficult, it is impossible. The previous section argues why authorial intentions cannot be proven in the objective, scientific sense. Secondly, the claim that the creator always has an *executive* intention is presumptuous. In many instances, the executive intention is that of a third party, such a religious institution that orders a painter to portray Christ in order to instill a sense of obedience in church-goers, or a wealthy internet client who commissions a digital artist to draw pictures of furries.

Additionally, Swirski calls anti-intentionalists hypocrites for invoking the author in their literary analysis despite claiming to discard him:

> Even the most ardent anti-intentionalists are, after all, inconsistent in their exegetical travails, referring to aims, authors, contexts, or aesthetic attributes as a matter of course—as they should, given that their theories bear little resemblance to what their contacts with literature are about. (139)

However, the elimination of objective authorial intent does not preclude the interpretation of the semantic (or otherwise) content of an artwork, which can still be a component of artistic appreciation. Rich subjective speculation about the meaning of a given art piece may still take place. Aesthetic reflection involves imagining what the creator(s) of an artwork could have felt and thought while making it. Audiences search for patterns, attribute motives, construct theories,

and play with ideas until they develop a viable interpretation of an artwork. This is an imaginative process that engenders aesthetic pleasure and prompts introspection. It is an interplay of detection and projection. The removal of the author need not remove the concept of "the authored text", as the next section will continue to argue.

1.4 Death of the Author

1.4.1 Killing the Author

The question of authorial intent can be considered on a more fundamental level by examining the concept authorship itself. In stark opposition to the intentionalist reverence of the author, Roland Barthes sees the contemporary idea of the "author" as a byproduct of enlightenment ideology, rather as an integral component of textual meaning:

> "In ethnographic societies the responsibility for a narrative is never assumed by a person but by a mediator, shaman or relator whose 'performance' - the mastery of the narrative code - may possibly be admired but never his 'genius'. The author is a modern figure, a product of our society insofar as, emerging from the Middle Ages with English empiricism, French rationalism and the personal faith of the Reformation, it discovered the prestige of the individual, of, as it is more nobly put, the 'human person'" (Barthes 142-143)

In contrast to traditional storytelling, the contemporary fixation on the individual author is a feature of western hegemony, not an intrinsic feature of art, Barthes argues. To support his claim that the (actual) author is not essential to the definition of art, I would invite the reader to recall the philosophical thought experiment known as the "Infinite monkey theorem", conceived of by the mathematician Émile Borel (189–196) . Imagine an infinite number of monkeys typing random words on an infinite number of typewriters. By way of logical necessity, one of those

monkeys is inadvertently going to write the complete works of Shakespeare. The person who would engage with the text would have no way of knowing whether the text was randomly-generated or written by Shakespeare himself. Readers retrieve (or project) an author from the text due to the historical context in which the plays' words and their meanings are immersed.

Foucault makes a similar observation, noting that "when the texts which we today call "literary" (narratives, stories, epics, tragedies, comedies) were accepted, put into circulation, and valorized without any question about the identity of their author." (109) Entire generations of people were able to understand the meanings of literary texts without presupposing the existence of authors. The texts had a life of their own, unfettered by the authoritative impositions of their creators. In the words of Jacques Derrida, there was (/is) "no outside-text" (158)

1.4.2 Convention

The Death of the Author prompts several questions. In a Foucauldian sense, if the generative power of the author were to be discredited, what other forces would replace its function and determine the meaning of artworks? Why is there such a high level of accord in the interpretation of art if there is no authority to definitively name the point of an artistic creation? An answer provided both by anti-intentionalists and proponents of the Death of the Author theory is convention:

> It is discovered through the semantics and syntax of a poem, through our habitual knowledge of the language, through grammars, dictionaries, and all the literature which is the source of dictionaries, in general through all that makes a language and culture (Wimsatt, Beardsley 477)

> "We know now that a text is not a line of words releasing a single 'theological' meaning (the 'message' of the Author-God) but a multi-dimensional space in which a variety of

> writings, none of them original, blend and crash. The text is a tissue of quotations drawn from the innumerable centers of culture...His only power is to mix writings, to counter the ones with the others, in such a way as never to rest on any one of them. Did he wish to express himself, he ought at least to know that the inner 'thing' he thinks to 'translate' is itself only a ready-formed dictionary, its words only explainable through other words, and so on indefinitely." (Barthes 146)

According to this view, meaning itself is a public domain. Individual authors rely on previously established linguistic norms and symbolic representations to produce their works. All that is within their power is to rearrange already-existing metaphors and words. This act of reconfiguration is not dissimilar to computers creating new works from databases. This comparison will be investigated more extensively in the next chapter (Manufacturing Intent). In any case, this conceptualization of the creative process as a web of interconnected meanings, works, and artists presents a threat to the enlightenment view of the individual genius author. All of his seemingly novel ideas are merely modifications of ideas that had been embedded into his cultural and linguistic realities by his predecessors. I argue that the postulation of such a rigid historical continuity implies determinism. Since the generation of "new" ideas was initiated by forces beyond the control of the author, she cannot claim complete ownership of them. Ideas are viral. Innovation is therefore depersonalized outside of an individualistic paradigm.

The conventional theory of pictorial representation demonstrates how meaning can be absorbed collectively. Artistic styles of dissimilar cultures represent objects in discrete, mutually-unintelligible manners. Much like mastering a language, one needs to learn what an artistic symbol represents within a culture (rather than what the author intendeed) to decipher the meaning of an artwork. Carroll chronicles the system of representation endemic to Ancient Egypt:

> "In the Egyptian system of representation, the nose of a figure is shown in profile while, simultaneously, the eye is represented frontally. This is why it is sometimes referred to as "the frontal eye" style. In a typical Renaissance painting, the eye and the nose are presented from a uniform angle of perception—from the same perspective: if the nose is in profile, so is the eye." (Carroll 39)

Under such a system of representation, the role of the author is merely to perform cognitive functions that enable the collective development of artistic movements. Much like an algorithm, the human artist receives (cultural, linguistic, and pictorial) inputs, which she subsequently subjects to cognitive processes associated with making art, after which she produces an output. Someone who would be lauded as a genius author under an individualistic paradigm is nothing more than a cog in a cultural machine if we were to kill the author and replace her with convention alone. However, there is another theory of art whose "meaning-makers" can replace the role of the individual author.

1.4.3 The Collective Unconscious

In the philosophy of art, naturalistic theories recognize a simple truth: our brains are hardwired to detect patterns. Our species evolved to discern facial expressions since we are the most complex social animals on the planet. It makes sense, then, that fans of oil paintings can experience a sense of existential panic just by staring at the distressed face depicted in Edvard Munch's *The Scream.*

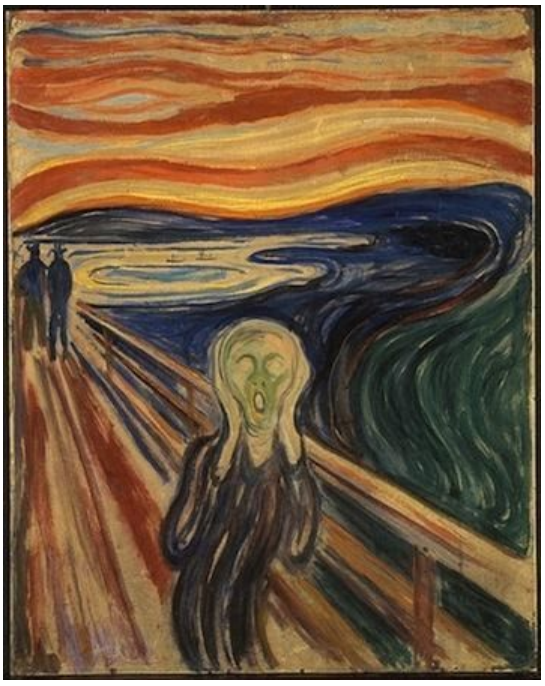

Figure 1: Edvard Munch, The Scream, 1893

One of the earliest articulations of our inherent capacity for collective meaning-detection came from Carl Jung, who pioneered the idea of the collective unconscious:

> "The existence of the collective unconscious means that individual consciousness is anything but a tabula rasa and is not immune to predetermining influences. On the contrary, it is in the highest degree influenced by inherited presuppositions, quite apart from the unavoidable influences exerted upon it by the environment. The collective unconscious comprises in itself the psychic life of our ancestors right back to the earliest beginnings. It is the matrix of all conscious psychic occurrences, and hence it exerts an influence that compromises the freedom of consciousness in the highest degree, since it is continually striving to lead all conscious processes back into the old paths." (Jung 112)

Jung posits a network of inter-connected symbols that is etched into our DNA -- a

comprehensive naturalistic account of why people across cultures share metaphors. For instance, the choice to use snakes as a proxy for danger in art is not random. Indeed, before our distant ancestors had descended from the trees into the tall grasses, it was critical to watch out for snakes slithering on branches. It is no surprise, then, that snakes epitomize dangers in human narratives, such as in the Biblical story of Adam and Eve, in which a talking serpent was employed to symbolize peril.

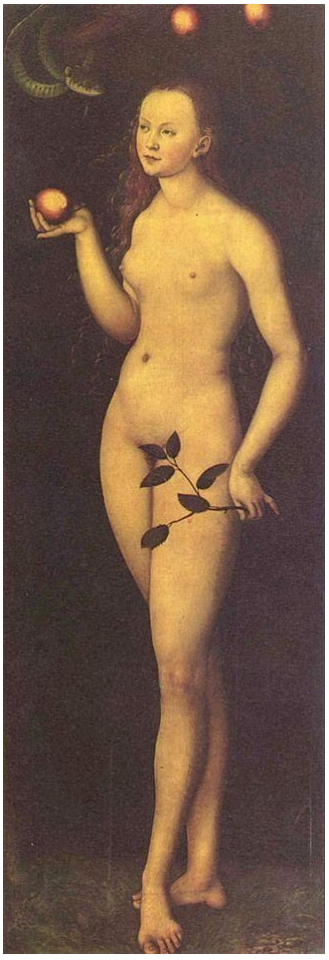

Figure 2: Lucas Cranach the Elder, Eva 1528

Given that human-made artworks constitute the data fed to machine learning algorithms, AI artists map our primordial intuitions onto their respective symbols. They learn our inherent biases. As will be demonstrated in a later section, this capacity for AI to replicate human prejudice will prove advantageous for the creation of computer-generated art. It is important to note that conventionalism and naturalism are not mutually exclusive. What's more, both of them

supplement the theory of hypothetical intentionalism which I will later support.

1.4.4 Authored Texts

The idea of an "authored text" was introduced by Michel Foucault, which was concisely summarized by Peter Lamarque: "All relevant claims about the relation between an authorasperson and a text are reducible to claims about an authoredtext." (326) Lamarque elegantly co-opts the idea of an 'authored text' to construct an argument against the death of the author. He points out how the dismissal of the 'actual' author does not oblige us to discard the idea of an 'authored text':

> Certainly the qualities of unity, expressiveness and creative imagination are still sought and valued in literary works, indeed they are bound up with the very conception of literature. If possession of these is sufficient for something's being an authored text, then authored texts are not dead. (324)

While Lamarque's paper would later challenge the necessity for the separation of the authored text and the actual author, I will merely appropriate the quote above to justify hypothetical intentionalism. An authored text has qualities (unity, expressiveness, creative imagination) that give the audience just cause to imagine what the author must have intended. This imaginative, constructive process should not to be mistaken for the investigative objective of actual intentionalism, which seeks to uncover the 'correct' intention ordained by an artwork's creator. Instead of positing an independent existence of the author outside of the text, the authored text shackles her to the artwork, rendering her existence outside of it obsolete. Therefore, external evidence of autorial intent becomes irrelevant. The only mechanism by which one can (re)construct authorial intent is by engaging with the work of art.

1.5 Hypothetical Intentionalism

1.5.1 Imaginary Intentions

The position in this discussion which I am defending is that of hypothetical intentionalism. I maintain that the author's intentions cannot be objectively uncovered but can be (re)imagined by the audience, who are prompted by the elements of a given artwork to arrive at certain conclusions about its meaning. The reason for the relatively high level of agreement among interpreters of art is that they share the cultural and biological backgrounds (conventionalism and naturalism, respectively) necessary to decode the meaning embedded in artworks.

This theory of intentionality is congruent with computer-generated art as it does not concern itself with the mental states of the real author, since they are not observable within the artwork itself. However, it is possible to observe aesthetic features that prompt the cognitive construction of a hypothetical author. In this paper, I will utilize arguments made by proponents of various variants of hypothetical intentionalism, taking from each what I consider their strongest aspects.

1.5.2 Utterance Meaning

William E. Tolhurst distinguishes between the meaning of a sequence of words and an utterance meaning, the latter of which is contingent upon the pragmatic context in which it is made (4). In his essay *On What a Text is and How It Means*, he defends the position that textual meaning is the same as utterance meaning. I agree with his position. Deftly, Tolhurst distances himself from actual intentionalists by demarcating the boundary between utterer's meaning and utterance meaning:

> Utterer's meaning and utterance mining are also distinct. Utterer's meaning is just

> whatever an utterer means by his use of a linguistic token. It might be thought that utterer's meaning is the same as utterance meaning, but the two must not be conflated. Although people on the whole do manage to write and say what they mean, die identity between what a person means by his utterance and what that utterance means is only contingent and not necessary. People sometimes fail to say what they mean: malapropisms and slips of the tongue are a part of everyone's linguistic experience. Thus utterer's meaning is not always utterance meaning. (4-5)

This way of understanding meaning helps my argumentation in two ways. Firstly, it disqualifies both extreme and moderate intentionalism from being useful in determining the meaning of an artwork due to the possibility of an author completely failing to match her utterance meaning to her utterer's meaning. Granted, moderate intentionalism merely claims that readers should value authorial intent from external evidence if that would yield a better interpretation than by deferring to convention. Nevertheless, since moderate intentionalism fails to establish a viable epistemic mechanism by which it determines what is the better or more correct interpretation, it does not escape the criticisms directed towards its more radical variant. Secondly, as I will write about at length later on, artificial neural networks with no internal mental states (that we know of) can generate art that can be interpreted by audiences *as if* there was an author. The idea that utterance meaning constitutes textual meaning compels audiences to interact with the art they are consuming; they are invited to reverse-engineer an author from the artwork. This imaginative process enriches the experience of art appreciation as well as rescues computer authors from intentionalist attacks.

1.5.3 *Departing from Levinson*

Jarrod Levinson is Tolhurt's successor when it comes to defending hypothetical intentionalism. His defences, on the other hand, modify Tolhurt's theory so he can more easily

evade rebuttal from its critics. In this section, I will engage with some of Levinson's ideas of hypothetical intentionalism so that (1) my arguments can be understood more clearly and so that (2) I make a meaningful contribution to the debate about hypothetical intentionalism. In his collection of essays *Aesthetic Pursuits*, Levinson makes the following arguments to which I will be responding, the first one being:

> Put otherwise, the view holds that it is a best hypothesis about authorial intent, and not authorial intent per se, that is constitutive or determinative of central literary meaning. (146)

Departing from Tolhurt and Levinson, I do not see the utility in claiming that it is necessary to hypothesise about the *best* possible authorial intent. Rather, I think there is descriptive utility in the theory when it explains why clusters of agreement form when it comes to interpreting a work. People who come from similar cultural and linguistic backgrounds tend to form similar hypotheses about authorial intent, which subsequently generate public meaning. I think it is not necessary to claim that there is such a thing as a "best hypothesis" or an "ideal audience", as (1) these things are difficult to metaphysically prove and (2) can be replaced with the more explanatory, modest proposal that certain audiences form certain clusters of hypotheses regarding authorial intent that go on to inform our collective understanding of a work's meaning.

Levinson makes another argument with which I disagree:

> the view I defend is not that the interpreter's task is to hypothesize an author, and subsequently, what such a hypothetical agent might have intended, but rather to hypothesize, in a fully contextually informed manner, about the actual author. (150)

While Levinson is in the minority when it comes to hypothesising the actual author's intention as opposed to a general author's intention, it is worth engaging with his assertion. The phrase "fully contextually informed manner" either implies information about the author that exists outside of the text (external evidence) or it does not. If it does, then it is a roundabout way of supporting modest intentionalism, which also takes both internal and external evidence to determine a fully contextually informed interpretation. If it does not imply familiarity with the author outside of the text, then imagining that specific author is exactly the same as imagining a general author, which makes his caveat redundant.

Therefore, the idea of a general, undefined author being hypothesized is one I support. Such an account works well with artificial intelligence since, if the audience were to consider external evidence, they would conclude that artificial intelligence had no intentions at all, which would diminish artistic appreciation. Imagining the possible mental states that *could have* led to the creation of computer-generated art, however, is a much more fruitful endeavor.

1.6 Conclusion

This overview of the authorial intent debate sought to expose the inadequacies of actual intentionalism. Moreover, analyzing how the public constructs meaning by engaging with art in culturally and linguistically informed settings demonstrates the advantages of hypothetical intentionalism. It provides a viable explanation of why groups of interpreters agree on a work's meaning, even if it departs from the author's intention. The nuances between different versions of hypothetical intentionalism provide room for further discussion.

1. Manufacturing Intent

2.1 Introduction

The second half of this paper will determine whether the conditions for authorial intent developed in the first half are compitable with the technology available today. An examination of what artificial intelligence can and cannot learn will illuminate to what extent computers can meaningfully produce art. The concept of verisimilitude will serve as a linguistic bridge between

the worlds of computer science and the philosophy of art.

2.2 Learning

In the introduction to the 30th-anniversary edition of his groundbreaking book on genetics *The Selfish Gene*, the biologist Richard Dawkins noted that "personifying genes, if done with due care and caution, often turns out to be the shortest route to rescuing a Darwinian theorist drowning in the muddle." (xii). This process of personification and mental-state attribution is critical to the theory hypothetical intentionalism. More importantly, it is critical to understanding computer-generated art.

My primary argument is that computer-generated art is art provided that humans who interpret it can attribute the same level of meaning to it as they would to human-generated art. If the two hypothetical authors that emerge are ascribed artistic intentions of equal value and complexity, then there is no meaningful difference between computer-generated and human-generated art. This section will walk the reader through the steps of how machines play the imitation game; how they learn to mimic -- and eventually outpace -- human intelligence.

2.2.1 Machine Learning

The computer-generated art that will be discussed in this paper is based on machine-learning algorithms. The following subsections will elucidate the general principles guiding these algorithms. It is important to note, however, that this is a philosophical paper, so the technological elaborations will be kept simple and brief. Critically, I will place the implications of machine learning within a discursively-relevant context by applying philosophical jargon to various technical phenomena.

The most popular university textbook on artificial intelligence *Artificial Intelligence: A Modern Approach* references the definition for machine learning as proposed by the Turing test, which is that machine intelligence is able to "to adapt to new circumstances and to detect and extrapolate patterns." (Russell, Norvig, 2) While there are different types of machine learning algorithms, most of them learn how to become better at a given task through trial and error (via

randomization followed by improvement), without being explicitly told which steps to take.

2.2.2 Learning Categories

Analytic philosophy is known for conceptual analysis. Broadly speaking, it seeks to uncover the necessary and sufficient conditions of a concept it is exploring. A concept, in this sense, is a category, a set consisting of its particular instantiations. The philosophy of art, for instance, seeks to detect (or maybe define) the necessary and sufficient conditions of art. However, there are other ways of conceptualizing concepts. Let us take a closer look at how artificial intelligence learns what a concept is.

"A core objective of a learner is to generalize from its experience" (Bishop, 2006) This quote, taken from the book *Pattern Recognition and Machine Learning* aptly describes how learners (both human and AI) come to mentally acquire categories. Artificial neural networks (ANN), -- which are capable of visually recognizing objects -- are not taught necessary and sufficient conditions to demarcate them. Instead, their nodes (artificial neurons) are taught to detect patterns just like human brains. To use philosophical vocabulary, they generalize through the process of induction. More specifically, they detect visual similarities in pictures (based on thousands of examples fed to their databases), which they then sort into categories. This process echoes theory of family resemblance:

> And the result of this examination is: we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail. I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way.— And I shall say: 'games' form a family. And for instance the kinds of number form a family in the same way. Why do we call something a "number"? Well, perhaps

> because it has a—direct—relationship with several things that have hitherto been called number; and this can be said to give it an indirect relationship to other things we call the same name. And we extend our concept of number as in spinning a thread we twist fibre on fibre. And the strength of the thread does not reside in the fact that some one fibre runs through its whole length, but in the overlap" (Wittgenstein 32, 33)

It would appear that Wittgenstein did only come up with a coherent account of how humans intuitively understand concepts, but he also predicted the way in which artificial intelligence would imitate human neural processes. Recognizing that machine learning algorithms can be trained to distinguish between different categories is the first step in bridging the gap between the capabilities of human and AI authors.

2.2.3. Learning Form and Style

Neo-formalism defines art in the following way:

> x is an artwork if and only if (1) x has content (2) x has form and (3) the form and the content of x are related to each other in a satisfyingly appropriate manner (Carroll 125)

Even philosophers who do not subscribe to neo-formalism as an ontological basis for art are likely to concede that form and style are significant elements of many artworks. It is worthwhile, therefore, to investigate the ways in which artificial intelligence is able to implement the two in its generation of art.

Scientists at the Werner Reichardt Centre for Integrative Neuroscience and Institute of Theoretical Physics, University of Tubingen, Germany trained a deep learning network to (1) learn styles and shapes and (2) combine them. In their paper, *A Neural Algorithm of Artistic Style*, they note: "in key areas of visual perception such as object and face recognition

near-human performance was recently demonstrated by a class of biologically inspired vision models called Deep Neural Networks" (Gatys, Ecker, Bethge 1). They employed Convolutional Neural Networks (CNN) to extract the style and content of a painting separately via thousands of layers (Gatys, Ecker, Bethge 2) -- hence the term "deep" learning. As a result, the neural network was able to adapt the content of the *Neckarfront* to numerous different styles.

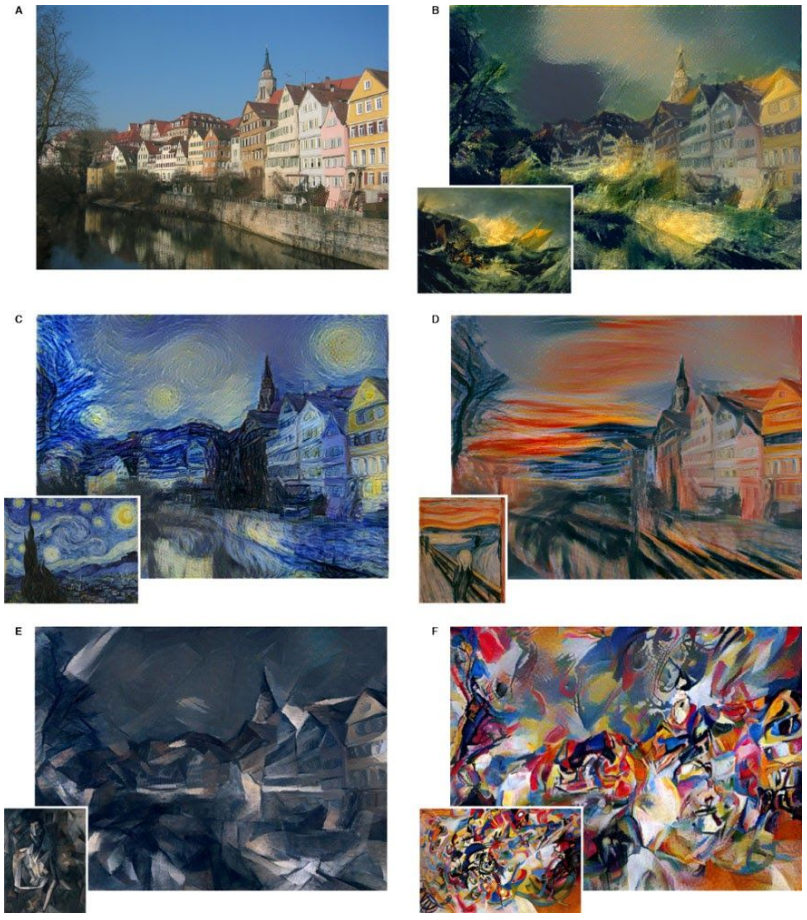

Figure 3: Neckarfronts, Convolutional Neural Network, 2015

With surgical precision, the neural network is able to depict the content o a picture X in the style of a picture Y, which meets the requirement of neo-formalism. One might object that the Convolutional Neural Network did not truly do any creative labor; rather it just rearranged previously-created elements in a harmonious fashion. It did not innovate, it copied. I would respond to such an objection by referring the critic back to the section on conventionalism, which argued that human authors do the same. Human artists exist within a historically-defined artistic practice that is modulated and constrained by culture, convention, and language. It is not

uncommon for human painters to contribute to a new style by borrowing existing tropes and forms from previous artistic movements and reconfiguring them. Human creativity, just like Convolutional Neural Networks consists of reconfiguration. What seems unique to humans is the creation of new styles, yet even that process can be examined historically; each new movement can be understood as a continuation of its predecessors. Perhaps CNNs can be taught how to distort already-existing elements significantly enough so as to usher in a new genre.

In any case, there is evidence that neural networks already resemble human brains: "Extracting correlations between neurons is a biologically plausible computation that is, for example, implemented by so-called complex cells in the primary visual system" (Gatys, Ecker, Bethge 9). Consequently, the critic of Convolutional Neural Network artists is either forced to make the concession that humans are never truly creative, seeing as our visual system rearranges patterns much like neural networks (hence we are the same) or assert that both humans and AI authors are capable of some form of innovation.

2.2.4 Learning Meaning

Presupposing that (1) authorial intent is a prerequisite for the metaphysical status of art and (2) that authorial intent is always inferred by the readers, artificial intelligence needs to be capable of producing works whose audiences will judge it to possess authorial intent. In the status quo, it can do just that, albeit very rarely. To increase the output of "meaningful" works, meta-authors may be of assistance.

The idea of a meta-author working in tandem with a computer to produce art was first introduced in *Gödel, Escher, Bach: an Eternal Golden Braid*, a speculative science book written in 1979 by Douglas Hofstadter. He posits that humans using artificial intelligence to engender art are "meta-authors" whose role in art-production is supplementary; they assist the real author -- the AI. An initial objection that comes to mind is that, in other mediums, artists

use tools (paintbrushes; drums; keyboards) to help them realize their vision, yet one would not call guitars and pencils “artists” and their users “meta-artists”. Hofstader explains the distinction with the following quote:

‘The distinction between author and meta-author is sharply pointed up in the case of computer composition of music. There are various levels of autonomy which a program may seem to have in the act of composition. One level is illustrated by a piece whose "meta-author" was Max Mathews of Bell Laboratories. He fed in the scores of the two marches "When Johnny Comes Marching Home" and "The British Grenadiers", and instructed the computer to make a new score--one which starts out as "Johnny", but slowly merges into "Grenadiers"’ (609)

The key distinction here is that the program does “creative” work by composing the music itself. Drums do not have the ability to compose music, neither do paintbrushes have the ability to conceptualize shapes. Therefore, programs are unique in their ability to make compositions, be they visual, aural, or narrative-based. Nevertheless, there is still space needed for a “meta-author” who feeds inputs into the program and selects outputs which he deems to have artistic value. Additionally, in accordance with the theory hyphothethical intentionalism, the meta-author could also serve the function of a meta-audience. To demonstrate the necessity of a meta-audience, let us consider the sale of *Edmond de Belamy*.

Created by a Generative Adversarial Network, the *Portrait of Edmond Belamy* was sold for $432,500 (Turnbull, *The Conversation*). The neural network was fed thousands of works by famous painters featuring human faces. First, the network was instructed to detect patterns in the paintings and form categoires of facial features. Then, it randomized the relevant data to generate new paintings. Finally, a network called “the discriminator” would determine whether the newly-generated painting resembled the originals. While the code was written by 19-year-old Robbie Barrat, the painting itself was printed by a trio of French students who call

themselves “Obvious”. They borrowed Barrat’s code to bring *Edmond de Belamy* into existence.

Figure 4: Obvious, Edmond de Belamy, 2018

The Generative Adverserial Network produced tens of thousands of paintings, many of which were unremarkable: they would lack structure and cohesion, failed to awaken aesthetic appreciation, or did not resemble human faces at all. Still, some of paintings were a relative success. Therein lies the need for a meta-audience. In accordance with the theory of hypothetical intentionalism, for a work to be considered an artwork, an audience must be prompted to attribute artistic intent to the work. However, thousands of the randomly-generated paintings did not meet this criterion. Hence, the “meta-authors” have to become “meta-audiences” in order to predict the success of an artwork. Plainly put, they need to hypothesize what real audiences would hypothesize what the authorial intention of the artwork was. Based on this meta-hypothesis, meta-audiences could select quality artworks from AI. What’s more, they could teach the AI to produce art of a higher quality. By rating certain paintings as “high in hypothetical meaning”, it is theoretically possible to train the AI to consistently print paintings

that humans consider to be imbued with authorial intent. Artificial selection works much like natural selection, but faster, meaning that AI has the potential to rapidly master the mimicry of meaning.

2.3 The Turing Test

2.3.1 The test

The Turing Test, proposed by Alan Turing (1950), was designed to provide a satisfactory operational definition of intelligence. A computer passes the test if a human interrogator, afterposing some written questions, cannot tell whether the written responses come from a person or from a computer. (Russell, Norvig 3)

It should go without saying that in 2020, computers are still not indistinguishable from humans. On the other hand, the Turing Test still provides a useful framework for determining whether computers can be indistinguishable from humans within a certain specialized domain. In some areas, the problem-solving capacities of computers are greater than that of humans, such as calculation, for instance. This section will argue that, in the status quo, some algorithms already manage to outperform humans in the industry of manufacturing intent.

2.3.2 Verisimilitude

A useful way to familiarize philosophers with the Turing Test is to reformulate it in Platonistic terms. Plato's disdain towards art is made apparent by his mimetic articulation of the aforementioned. In the *Republic* (596c–e), he compares art to holding a mirror and reflecting the outside world with in. According to him, art is imitation. Not only is it an imitation of material objects, it is a third-order imitation, since material objects themselves are imitations of eternal Platonistic ideals. Verisimilitude refers to the capacity of an artwork to accurately imitate the thing it is representing.

Such a notion has utility in describing the Turing Test: verisimilitude is an indicator of how human-like the AI is. More specifically, when it comes to art, computer-generated works fit

neatly into the Platonistic hierarchy -- they are fourth-order imitations. Computed-generated works do not seek to imitate eternal ideals, nor do they seek to imitate material object -- their exact purpose is to imitate art. As mentioned before, they scan databases of artworks to obtain an understanding of what they are and what they depict. By using "verisimilitude" in this context, philosophers are given a metric by which they can measure how effectively the artistic process is mimicked by AI. The Goal of the AI, under the paradigm of hypothetical intentionalism, is to achieve perfect verisimilitude of authorial intent.

Let us take a look at another member of the Belamy family to see how this process plays out.

Figure 5: Obvious, *Madam de Belamy*, 2018

Madam de Belamy was created by an entity with (presumably) no internal mental states. However, it could be argued that the reason why the painting was selected for the auction out of thousands of other paintings was its capacity to prompt authorial intent speculation. Plausibly, one could hypothesize that this painting was intended to depict depression. The madam is either looking down or her eyes are closed -- it is difficult to tell due to the blur, which in and of itself is a feature of expressionism, which focuses of affect. Additionally, the blue eyeshadow

complements the rest of her outfit which, taken together, could symbolize a cold personality or emotional detachedness. Judging from her fashion sense, it is evident that the woman is upper-class, which might be why, even when frowning, she tries to maintain a dignified expression. All in all, I conclude the author must have intended the painting to represent the repression of negative emotions. Although, naturally, I am only one biased person. Therefore, I cannot determine the verisimilitude of authorial intent in this picture alone. *Madam de Belamy* could warrant further empirical research on the effectiveness of its mimicry pertaining to authorial intent by asking study participants to rate its expressiveness. Regrettably, that is outside the scope of this paper.

2.3.3 Artificial Selection

Charles Darwin's theory of evolution is the cornerstone of modern science. Evolutionary biology is predicated upon the fact that every identifiable trait in a living being has been naturally selected. A much faster form of trait selection is emerging: artificial selection. This term typically refers to humans breeding species of animal and plants, but it is also applicable to machine learning.

In 2016, an artificial intelligence (AlphaGo) beat the world champion in Go, a board game so complex that it has more move combinations that atoms in the universe (BBC, 2016). Fascinatingly, it was not instructed *how* to play the game. Instead, it was a given goals: prioritize winning and avoid losing. The deep learning network played millions of games against itself, trying out random moves until it determined which sequences helped its goals of prioritizing winning and avoiding losing. This formula of randomization *plus* selection for a goal (be it survival or winning) is identical in natural selection and artificial selection, the latter of which is astonishingly faster. Is it possible, then, as hinted in the previous chapters, for AI to self-select for meaning? Yes.

Academics from the university of Rutgers working on the research paper *CAN:*

Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms had participants give an answer ranging from 1-5 to the following statements concerning artworks created by humans versus artworks created by machines:

> Q1: As I interact with this painting, I start to see the artist's intentionality: it looks like it was composed very intentionally.
> Q2: As I interact with this painting, I start to see a structure emerging.
> Q3: Communication: As I interact with this painting, I feel that it is communicating with me.
> Q4: Inspiration: As I interact with this painting, I feel inspired and elevated. (Elgammal, Liu, Elhoseiny, Mazzone 17)

Against their expectations, the neural network outperformed human artists in their ability to produce intention-attribution in the minds of the human subjects:

> We also hypothesized that human subjects would rate art by real artists higher on these scales than those generated by the proposed system. To our surprise the results showed that our hypothesis is not true! Human subjects rated the images generated by the proposed system higher than those created by real artists, whether in the Abstract Expressionism set or in the Art Basel set. (Elgammal, Liu, Elhoseiny, Mazzone 18)

While this may be only one survey, it is a promising find in the field of artificial intelligence. The capacity to create an illusion of authorial intent is a trait for which artificial intelligence can actively select. Authorial intention, it seems, is a product that can be manufactured.

2.4 Conclusion

The rapid technological development of artificial intelligence has resulted in an

explosion of computer-generated art that meets the ontological criteria defended in the previous chapter. Via selective imitation and reconfiguration, neural networks can learn to how make audiences believe that randomly-generated art had artistic intent behind it. This revelation opens up numerous philosophical considerations about the implications and applications of computer authorship.

References

Barthes, Roland, and Stephen Heath. I*mage, Music, Text*. New York: Hill and Wang, 1977. Print.

Carroll, Noël. *Philosophy of Art: A Contemporary Introduction*. London: Routledge, 2010. Print.

Cauer, Wilhelm; T*heorie der linearen Wechselstromschaltungen*, Vol.I, Akademische Verlags-Gesellschaft Becker und Erler, Leipzig, 1941.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning* (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Dawkins, Richard, 1941-. *The Selfish Gene*. Oxford ; New York :Oxford University Press, 1989.

Elgammal, Ahmed & Liu, Bingchen & Elhoseiny, Mohamed & Mazzone, Marian. (2017). CAN: *Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms.*

Gallop, Jane. *The Deaths of the Author: Reading and Writing in Time*. Duke University Press, 2011. JSTOR, www.jstor.org/stable/j.ctv11cw7dr. Accessed 13 Sept. 2020.

Gatys, L. A., Ecker, A. S. & Bethge, M. (2015). *A Neural Algorithm of Artistic Style* (cite arxiv:1508.06576)

Heller, Reinhold. Edvard Munch: The *Scream.* [New York] :[Viking Press], 19731972.

Hirsch, E. D. *Validity in Interpretation*. Yale University Press, 1967. JSTOR, www.jstor.org/stable/j.ctt32bd9k. Accessed 13 Sept. 2020.

Hofstadter, Douglas R., 1945-. *Gödel, Escher, Bach : An Eternal Golden Braid*. New York

:Basic Books, 1979.

Jung, C. G. (Carl Gustav), 1875-1961. T*he Archetypes and the Collective Unconscious.* [Princeton, N.J.] :Princeton University Press, 1980.

Lamarque, P. V. (1990). *The Death of the Author: an Analytical Autopsy. British journal of aesthetics*, 30, 319-331.

Levinson, Jerrold. *Aesthetic Pursuits: Essays in Philosophy of Art*. Oxford: Oxford University Press, 2016. Internet resource.

Michel Foucault, *Authorship: What is an Author?*, Screen, Volume 20, Issue 1, Spring 1979, Pages 13–34,

Plato. *Plato's The Republic.* New York :Books, Inc., 1943.

Russell, Stuart J, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, N.J: Prentice Hall, 1995. Print.

Swirski, Peter. *Literature, Analytically Speaking.* University of Texas Press, 2010. JSTOR, www.jstor.org/stable/10.7560/721784. Accessed 13 Sept. 2020.

Tolhurst, William E. *On What a Text is and What It Means The British Journal of Aesthetics*, Volume 19, Issue 1, 1979, Pages 3–14, https://doi.org/10.1093/bjaesthetics/19.1.3

Turnbull, Amanda, "*The Price of AI Art: Has the Bubble Burst?"* The Conversation, 12 Sept. 2020, theconversation.com/the-price-of-ai-art-has-the-bubble-burst-128698.

Unknown, BBC News*, Artificial Intelligence: Google's AlphaGo Beats Go Master Lee Se-Do*l.", BBC, 12 Mar. 2016, www.bbc.com/news/technology-35785875.

Weisberg, J. (2015). H*ard Problem of Consciousness.* [online] Iep.utm.edu. Available at: http://www.iep.utm.edu/hard-con/#H2 [Accessed 28 Aug. 2020].

Wimsatt, W. K., and M. C. Beardsley. *"The Intentional Fallacy*." The Sewanee Review, vol. 54,

no. 3, 1946, pp. 468–488. JSTOR, www.jstor.org/stable/27537676. Accessed 13 Sept. 2020.

Wittgenstein, Ludwig, and G E. M. Anscombe. *Philosophical Investigation*s. Oxford, UK: Blackwell, 1997. Print.

Émile Borel (1913). "*Mécanique Statistique et Irréversibilité*". J. Phys. (Paris). Series 5. 3: 189–196. Archived from the original on 2015-11-30. Retrieved 2019-03-23. (The journal appears to not be archived back to 1913)