The Hermeneutics of Computer-Generated Texts

Abstract

As cultural circumstances become increasingly digital, the importance of theoretical frameworks guiding calculated considerations of authorial intention and reader response is being reaffirmed. The framework proposed in this article is that of hermeneutics: the study of understanding, of processes of meaning-making. Although explicit application of hermeneutics has fallen out of fashion, the field is especially valuable for critically approaching digital texts. This article thus serves as a re-introduction to hermeneutics, particularly for digital textual study. It offers an overview of historical hermeneutical views, and then applies a hermeneutics perspective to a new kind of text made possible by digital technologies: computer-generated prose. Through analysis and repurposing of OpenAI's GPT-2 software, this paper argues that the reintegration of hermeneutics in digital textual studies may contribute to more comprehensive understandings of both human and computer intention, especially in instances of computer-generated texts. Digital technologies are changing conventional understandings of authorship and reader responsibility; hermeneutics helps us understand what these changes are, how they have come to be, and why they matter.

Introduction

In an article for the *Los Angeles Review of Books*, Professor John Farrell (2019) questions 'Why Literature Professors Turned Against Authors – Or Did They?' He opens with the assertion that:

> Since the 1940s among professors of literature, attributing significance to authors' intentions has been taboo and *déclassé*. The phrase *literary work*, which implies a worker, has been replaced in scholarly practice – and in the classroom – by the clean, crisp syllable *text*, referring to nothing more than simple words on a page. Since these are all we have access to, the argument goes, speculations about what the author meant can only be a distraction.

By the end of Farrell's article, though, it is clear that efforts to disregard authorial intention in acts of textual interpretation are fruitless and futile. 'Fortunately or unfortunately,' Farrell contends, 'it is impossible to get rid of authors entirely because the signs that constitute language are arbitrarily chosen and have no significance apart from their use.' The age-old dispute between authorial intention and reader interpretation continues, arguably becoming even more relevant as digital circumstances facilitate authorial collaboration (e.g. Wikipedia, Google Docs), as well as interactive experiences wherein readers are permitted greater (perceived) control over narrative progression (e.g. by following links in hypertext).

The study of hermeneutics is especially relevant for considering disputes about authorial intention and reader responsibility. In his 1969 book on the subject, Richard Palmer defines hermeneutics as the study of understanding. More particularly, '[i]t tries to hold together two areas of understanding theory: the question of what is involved in the event of understanding a

text, and the question of what understanding itself is' (10). This definition is, of course, intentionally broad to encompass a field that is replete with contradicting perspectives of (1) what understanding actually means and (2) how to foster such understanding. Hermeneutics is, in essence, the study of meaning-making processes, of individual and collective acts of interpretation. As Jens Zimmerman has observed in his *Very Short Introduction* (2015), the focus of hermeneutics is generally on the receiver, and how that receiver comes to make meaning from a message. However, hermeneutics may also encompass a message sender's efforts to be understood. For textual studies, hermeneutics is invaluable. How do we come to understand what we are reading? What does it mean to understand a text? What do our approaches to understanding say about the value of literary texts in our respective cultural contexts? From the 1960s to 1990s, hermeneutics was named as part of discussions about authorial intention, reader responsibility, and literary analysis; explicit recognition of the field has, however, fallen out of fashion. As Johanna Drucker observed in 2002 (686), 'poets, critics, scholars, teachers - all regularly and frequently overlook the bibliographical, graphical, and materially semiotic codes of the media of the printed texts on which they depend.' This paper aims to draw attention back to those semiotic codes by bringing hermeneutics once again to the forefront of literary analysis, highlighting how understanding processes of meaning-making, as well as paratextual and cultural elements, may contribute to deeper appreciations of textual value. It pays special attention to the implications of an increasingly digital cultural context that facilitates varied approaches to both reading and writing. Specifically, it questions whether current hermeneutic theories are sufficient in the advent of natural language generation (NLG) techniques that challenge conventional notions of authorial intention, readerly perceptions of authorship and interpretation, and potential intrinsic limitations of language used in hermeneutic

analyses. At the same time, it aims to diversity discussions of hermeneutics through inclusion of relevant scholarship by disciplinary groups that have largely been marginalised in this area.

There is no consensus amongst scholars about what constitutes a ‘good’, ‘correct’, or ‘true’ hermeneutics, and it is not the place of this paper to argue for one. Instead, we offer an abridged survey of prominent hermeneutical stances that literary scholars may wish to consider as they conduct their textual analyses. We do so to streamline the convoluted conversations related to hermeneutics, providing a fresh starting point for researchers who may wish to delve deeper into the realm(s) of understanding but are unsure of where to begin their journeys. We do not present theories in chronological order, but group similar perspectives for ease of reference. Following the survey, we present our own perspectives on where hermeneutics fits within literary study in digital contexts. In particular, we consider how hermeneutics scholarship may guide interpretations of authorial intention and reader responsibility pertaining to computer-generated texts, whose authorship may be uncertain or obscured. From all of this, though, no real conclusion is reached – just deeper understanding of understanding.

The Emergence of Hermeneutics

While some histories of hermeneutics cite Aristotle’s *De Interpretatione* (*On Interpretation*) as the field’s origin, it was not until the Christian world adopted hermeneutics as its own study focused on biblical interpretation that the field took root. For hundreds of years, hermeneutics was primarily considered a theological discipline: a search for understanding the written scriptures to establish a closer relationship with God. Some scholars (Moules, 2002) hold that the term hermeneutics even stems from the name of the Greek God Hermes, who served as a

messenger. In her book about theological hermeneutics specifically, Marion Grau (2014: 80) explicitly argues for 'putting Hermes back into hermeneutics', advocating that 'the concept of hermeneutics be stretched to imply patterns of interpretive discursivity across religions and cultures. In fact, tracing the figure of Hermes and its characteristics may help us reframe the work and scope of theological hermeneutics by affirming polyvalence and indeterminacy as at the core of the work of interpretation.' Hermeneutics has long been associated with the sacred, with the reception of divine truths passed from deity to human – however polyvalent and indeterminate interpretations of these truths may be. For textual analysis, this theological perspective serves as a metaphor for truth-seeking in communicative processes. The writer is God, omniscient and omnipotent; the reader is human, flawed but seeking transcendence.

It was Friedrich Schleiermacher who expanded hermeneutics' scope to include virtually any written text, and it is for this reason that he is widely considered the father of modern hermeneutics. In his 1838 *Hermeneutics and Criticism and Other Writings* (first published as *Hermeneutik und Kritik mit besonderer Beziehung auf das Neue Testament*), Schleiermacher describes the process of textual interpretation as circular: this view has since been translated as the 'hermeneutic circle' (a term that was supposedly later coined by Heidegger). For Schleiermacher, '[t]he goal of hermeneutics is understanding in the highest sense. […] To this also belongs understanding the writer better than he understands himself' (1998: 228). This 'understanding in the highest sense' is implied to be the author's intention, which has been manifest through written language that reflects logics and ontologies associated with authorial intention. For religious texts, this author could be God; otherwise, the author is a human individual presumed to have unique thought. To achieve this understanding, the reader must consider both the author's language and thought. The former is 'the embodiment of everything that can be thought in it', as language is seen as constraining the expression of thought by

seemingly objective definitions of words (Schleiermacher, 1998: 229). The latter refers to the intention of the writer as it has been influenced by his subjective experiences, themselves influenced by 'the area of his time, of his education, and of his occupation – also of his dialect – where and to the extent to which this difference occurs in educated discourse' (Schleiermacher, 1998: 31). Experiences of language are not isolated to the subjective experiences of individuals, but are always social. The writer uses her own vocabulary to construct sentences that are comprehensible in light of the contemporary context, thereby contributing to that context, while the context itself constrains the writer's use of language. Hypothetically, the reader could 'understand the writer better than he understands himself' because the reader may consider textual output impartially, with one eye directed towards authorial intention and the other directed towards the cultural contexts of production and reception. We noted above that father of modern hermeneutics Friedrich Schleiermacher held that a text may have one correct interpretation: that intended by the text's writer. Paradoxically, though, such understanding comes from not only considering the textual output of any one individual, but also from considering the social contexts of textual production. In this way, the hermeneutic circle substantiates its name: the individual writer writes in light of personal and subjective experiences within a particular culture; in writing about that culture, the culture is reinforced, altered, and perpetuated.

Richard Palmer (1969: 87) offers a clear definition of the hermeneutic circle, which he refers to as the 'hermeneutical circle':

> Understanding is a basically referential operation; we understand something by comparing it to something we already know. What we understand forms itself into systematic unities, or circles made up of parts. The circle as a whole defines the

> individual part, and the parts together form the circle. A whole sentence, for instance, is a unity. We understand the meaning of an individual word by seeing it in reference to the whole of the sentence; and reciprocally, the sentence's meaning as a whole is dependent on the meaning of individual words. By extension, an individual concept derives its meaning from a context or horizon within which it stands; yet the horizon is made up of the very elements to which it gives meaning. By dialectical interaction between the whole and the part, each gives the other meaning; understanding is circular, then. Because within this 'circle' the meaning comes to stand, we call this the 'hermeneutical circle.'

In Palmer's definition, which he uses to summarise Schleiermacher's view, the hermeneutic circle not only assumes linguistic understanding, but also an understanding of the subject being discussed that is shared between the author and the reader. Yet the hermeneutic circle encompasses more than just recurrent processes. For Heidegger (1996: 143), the hermeneutic circle situates individuals within greater contexts of being; it is an interpretive strategy for characterising phenomenological experience. Heidegger's student Hans-Georg Gadamer (1988: 68) elaborates upon Heidegger's notion of multiple spheres of understanding, noting that the hermeneutic circle may also refer to concentric circles – circles of different sizes, sharing a centre – that contribute to a more harmonious representation of a text's parts and whole that may lead to its 'correct understanding'. In this way, the hermeneutic circle is swelling and shrinking. To use the words of foundational philosopher of hermeneutics Wilhelm Dilthey (1996: 231), '[a] whole should be understood on the basis of the particular and the particular on the basis of the whole. This contradiction generates the procedure of the hermeneut. He operates

with hypotheses.' At the centre of the circle sits the writer; the circle shrinks to focus on this individual, and gradually swells to encompass ever more.

Questions of Authorial Intention

For Gadamer (1988: 69), though, the 'correct understanding' of a text may not be the initial intention of the author when writing, but may actually be 'a participation in shared meaning' between the author and reader as the reader negotiates the text according to an expectation of the text as a manifestation of communicational intention. Further, there is what Gadamer (1988: 74-75) calls the 'anticipation of perfection'. He explains:

> Just as the addressee of a letter understands the news he receives and, to begin with, sees things with the eyes of the letter-writer, i.e., takes what the writer says to be true – instead of, say, trying to understand the writer's opinion as such – so we too understand the texts which are handed down on the basis of expectations of meaning drawn from our own relationship to the issues under discussion. And just as we believe the reports of a correspondent because he was there or in some other way knows better, so too we are basically open to the possibility that the text which has come down to us knows better than our own pre-opinion wants to admit.

Some theorists – aptly called intentionalists – adhere to Schleiermacher's understanding of a text as having a 'correct' interpretation, with a text's correct interpretation almost always being considered the intention of its writer. E. D. Hirsch (1967: 216) is perhaps the most adamant of these theorists, holding that '[t]his permanent meaning [of a text] is, and can be, nothing other than the author's meaning.' Hirsch (1967: 236) recognises the impossibility of readers ever

knowing with certainty that their interpretations of a text are correct, dismissing this issue with a statement that '[t]he interpreter's goal is simply this – to show that a given reading is more probable than others.' More recent papers support this argument (Jiang, 2018). For a broad example of this view, however, one need only look to the field of genetic criticism (*critique génétique*), wherein textual scholars analyse the preparatory material (e.g. manuscripts and letters) informing a text's final form. While not necessarily touting such intentionalist views as Hirsch, the practice of genetic criticism relies upon an implicit understanding of the text as a work of ever-developing authorial intention, as consideration of the authorial process being perceived as a means for deeper understanding of a text. Genetic criticism also implies a correct interpretation: that which was, at some time, intended by the author.

For opposing theorists, the author is considered altogether irrelevant for textual interpretation. This is what W. K. Wimsatt and M. C. Beardsley (1946) call the 'intentional fallacy': the understanding that the authorial intention driving a text's production is the 'correct' interpretation of that text. For Wimsatt and Beardsley, the author's intention while producing a text is subordinate – and perhaps even irrelevant – to the reader's experience of that text. The author cannot be resurrected through textual analysis; it is the text that is the primary source of meaning rather than any biographical details of the author that may have informed that text's production. Literary critics Roland Barthes and Michel Foucault similarly portray the author as an ideological product of a post-industrial culture of individualisation. Barthes (1977: 143), for example, describes authorship as 'the epitome and culmination of capitalist ideology, which has attached the greatest importance to the "person" of the author'. In Barthes' view, overemphasis on the biography of a text's writer to discern meaning from that text detracts attention from the complex social networks related to its production, dissemination, and reception, all of which Barthes argues are contained within the reader. He stresses authorship as a discursive function

rather than any particular individual. The ‘person’ of the author is more a cultural construct than a body of flesh. Likewise, Foucault explores the ways in which authorship is a cultural construct, a discursive function over any one person. Foucault (1998: 211-212) notes, for example, how the judicial system enforcing ownership rights, emerging around the turn of the nineteenth century, has played a vital role in establishing the ‘author function’ as we now know it. He stresses that such a system emerged to combat religious and social transgression; authorial attribution held writers accountable for their words. In this sense, authorial attribution recognized the individual in an effort to maintain a sense of order within society. Hence while modern reading practices are often regarded as solitary and individualistic, emphasising the subjectivity of each reader, they are nevertheless informed by social institutions such as legal claims to textual ownership (Ong, 1982; repr. 1997). Social circumstances thus inform not only the reading experience, but also the readers’ perceptions of who – or what – the author is. In these scholars’ views, the author is not so much an individual writer as an imagined figure, a product of contemporary contexts. Moreover, reading is a personal experience that occurs within a greater cultural context that informs readerly interpretation. Readers may identify with the events described in a work, or with the feelings conveyed, despite experiential distance. For those with this formalist view – now known as New Criticism – the purpose of literature is to explore the self, to reflect upon the personal, to arouse that which already exists within the reader.

Reception theory is a form of hermeneutics that explores the role of the reader in particular. Literary theorist Terry Eagleton (1996: 66) explains the fundamental premise of reception theory as being that of readers making conceptual connections between the text and the world within which the text exists. ‘The text itself is really no more than a series of “cues” to the reader, invitations to construct a piece of language into meaning,’ Eagleton writes. ‘Without

this continuous active participation on the reader's part, there would be no literary work at all.' Returning to the theological slant of natal hermeneutics, Gillian Beer (2014: 1) observes that 'human authors are much more like Satan than God: they can start things but not control consequences. Once the book is published goes out of their control, into the anonymous realm of the reader.' Reception theory considers reading to be an active process of constant interpretation, aligning well with the concept of the hermeneutic circle. Shifting between the particular vocabulary used in the text and one's more subjective personal experiences and frames of reference, the reader oscillates between considering the parts of the text – the words the author has used to 'cue' the 'concretisation' – and the text as a fundamentally social artefact. Gaps in the text are filled by readers' own understandings of the world around them or, perhaps for the more historically-minded readers, by readers' own understandings of the cultural contexts within which the text was produced.

For reception theorists, the reader is considered of more value to textual interpretation than the author. Not only do words have both literal and metaphorical meanings that may be distinct (even, at times, contradictory), but readers have entirely unique experiences of reading those words. In this way, the author is of little importance, and it is the reader whose experience of the text is considered foremost. Wolfgang Iser (1989) analogises a fictional text – but any text, really – as a playground in which an author plays games with readers through intentional acts of referring to and intervening in an existing world. The existing world to which Iser refers contains the sociological reality of textual production and reception; the fictional world exists only within the text. It is the reader's responsibility to negotiate these worlds through acts of meaning-making; while the creation of a text may have been driven by authorial intention, the text is interpreted by each reader as per that reader's unique perspective. The reader's understanding of the fictional world is informed by the existing world's realities. Stanley Fish

(1980: 45) recognises, however, that readers of the same language share internalised understandings of syntactic and semantic rules permitting mutual understandings, '[a]nd insofar as these rules are constraints on production – establishing boundaries within which utterances are labeled "normal," "deviant," "impossible," and so on – they will also be constraints on the range, and even the direction, of response'. Thus, while reading experiences may be highly individualised, interpretation is still informed by sociocultural circumstances that inform collective understandings of language, contributing to formation and maintenance of 'interpretive communities'. Even for Fish, though, the author and reader collaborate in their interpretive strategies. 'Readers who perform in the ways I have been describing', Fish (2001: 37) elaborates, 'do not ride roughshod over an author's intention; rather they *match* it by going about their business at once constrained and enabled by the same history that burdens and energizes those whom they read.'

There are myriad more perspectives that could be integrated into a survey of this kind. Underlying all of these approaches to textual hermeneutics, though, is the acceptance of an author-reader relationship. Regardless of whether or not a theorist holds that there is one 'correct' interpretation of a text, there is a more general assumption of both author and reader involvement in processes of meaning-making. This author-reader relationship is perhaps considered so obvious as to be omitted from explication, but it continues to be worthy of deeper consideration.

There are, of course, gaps in what has been relayed above. For one, theories of hermeneutics tend towards an assumption that the author's thoughts are almost entirely linguistic. If one favours authorial intention, how is one to accommodate the author's process of transforming thoughts into words? Language is in such instances a finite encoding of thought

that is limited by constraints of culture. Returning to Schleiermacher's assertion (1998: 228) that hermeneutics has an aim of 'understanding the writer better than he understands himself,' how can readers come to understand thoughts better than the mind that originated them through language's limited proxy of thought? A text is a surrogate of authorial thought, but not all thoughts unfold linguistically.

Applying Hermeneutics to Computer-Generated Texts

Indeed, not all works employ text as a communicative preference. The digital sphere is riddled with imagery, in static (photos) or moving forms (GIFs and videos). Even those works that are text-based take on markedly new forms in light of digital evolutions. For example, hypertextual and multimedia experiences fundamentally alter readers' experiences of any one text (Ensslin, 2020). Moreover, digital technologies allow for altogether new kinds of text that may bring conventional understandings of authorial intention and reader responsibility into question. This is the case not just for works of electronic literature (texts created digitally, for digital consumption), but also for more 'traditional' print books, as Jessica Pressman (2018) shows. Digital technologies change conceptions of authorship, and in turn change conceptions of the conventional author-reader relationship. Our survey began with an introduction to hermeneutics as being rooted in theology and the search for truth. As has been shown, modern theories of hermeneutics remain impregnated with this religious history. The role of God – the writer – has

changed, as has the role of the fallible human – the reader – but the search for truth ultimately continues. How, though, does this search continue in the realm of the digital?

Digital contexts accentuate both the benefits and shortcomings of hermeneutics as a field. For brevity's sake, we focus on one kind of text resulting from digital developments: computer-generated texts, whose 'authors' may not be readily apparent. Who – or what – do readers anticipate as the author of the computer-generated text? How do readers make meaning from these texts? Questions like these highlight the continued importance of hermeneutics as a fundamental part of textual studies. Authorship is always a collaborative endeavour (McGann, 1991). However, repositioning human roles in text production and reception prompts reconsideration of how meaning may be discerned from text. Situating discussions of this sort within their extended history of hermeneutics allows for richer analyses that appreciate the complexity of author-reader relationships in varied forms. Computer-generated texts make clear that it is not enough to only consider authorial intention. Perhaps these texts necessitate another kind of approach: one that includes beyond-intention meaning. Such an approach recognises that readers discern meaning beyond that intended by the author, while at the same time recognising the value of authorial intention.

Computer-generated texts abound, and many are altogether indistinguishable from human-written texts. Indeed, developers and critics of natural language generation (NLG) systems, typically considered a branch of artificial intelligence (AI), often adhere to a 'Turing test' standard of evaluation. If a reader is unable to identify the text in question as computer-generated, the text has passed the Turing test. Passing the Turing test is sometimes seen as proof that an algorithm or AI has reached human-level intelligence. However, some authors like Janelle Shane (2019: I. 525) believe that 'the test isn't actually a good measure of

algorithmic intelligence' because it's all too easy to pass it if one 'makes the topic [of the text] narrow enough'. In general, 'the narrower the AI, the smarter it seems'. Martin Eve reflects upon a recurrent neural network (RNN) he trained to generate textual output that mimics the writing style of articles from the literary studies journal *Textual Practice*. He (2017: 46) explains that 'the network has no motivation towards communication and no epistemological goal except to achieve ever more perfection in its stylistic mimicry of the articles in *Textual Practice*.' Despite the RNN producing text that is understandable, albeit verbose, Eve is unwilling to call the system 'author'. He denounces the proliferation of NLG systems as 'alarming' not only because these systems threaten to mechanise current labour practices, but also because he believes that they depreciate narrative quality and undermine the work of narrative scholars. According to Eve (2017: 51):

> [I]n achieving a mimesis of human writing – remember, a measure of intelligence formed only by anthropocentric reference to the human – computational writing asks us to imagine a world in which there are no more humans undertaking such labour. Such thinking only emerges, though, in the imagined substitution of the human with human-like automata. This imagined world is both a post-anthropocentric world for writers and a world in which a writing machine that is legitimated by human-like characteristics is inscribed at the centre. It is concurrently a world in which we have no benchmark of contemporary writing success, but one that is nonetheless dominated by machines that meet that nostalgic target.

Eve's hypothetical world exemplifies an approach to thinking about computer-generated texts that tends towards apocalyptic outcomes of human labour because computers can produce adequate imitations of human-written texts. Similar disparaging views were expressed by the

ordinary readers asked for their opinions about computer-generated texts; many concerns related to jobs – especially ‘creative’ jobs like journalism and fiction writing – being dominated by NLG systems capable of mass-producing reasonable-quality material at unprecedented rates (Henrickson, 2019). Evidently, there is not only reluctance to assign authorship to a computer, as per Eve, but reluctance to accept the technology more generally.

One of the most public affirmations of this reluctance was in 2018, when artificial intelligence research company OpenAI made headlines with their decision not to release their newly-developed text generator GPT-2. ‘New AI fake text generator may be too dangerous to release, say creators’, the title of one news article read in February 2019 (Hern); ‘New AI Development So Advanced It's Too Dangerous To Release, Says Scientists’, another title read (Martin, 2019). OpenAI had supposedly developed a text generator capable of producing syntactically, semantically, and stylistically impressive texts: texts so impressive that they could potentially lead to the proliferation of convincing computer-generated ‘fake news’. An abridged example of GPT-2’s output follows. Note that GPT-2 generates its texts in response to human-inputted prompts (‘Better Language Models [Sample 2]’, 2019):

> System prompt (human-written): *A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.*
>
> Model completion (machine-written, first try): The incident occurred on the downtown train line, which runs from Covington and Ashland stations.
>
> In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

> ‘The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,’ said Tom Hicks, the U.S. Energy Secretary, in a statement. ‘Our top priority is to secure the theft and ensure it doesn’t happen again.’

OpenAI’s choice not to release GPT-2 was likely motivated by the desire for publicity, but the discussions spurred by this choice highlight questions relevant to a consideration of hermeneutics for modern literary analysis. Do these texts render authorship obsolete? How are readers to make meaning from a text that is not produced by one intention-driven human, but by a statistical mind that composites all the authors from which it has learned?

To demonstrate the applicability of hermeneutics scholarship to such instances, let us return to the hermeneutic circle, which Richard Palmer (1969: 87) describes as a means for ‘understand[ing] something by comparing it to something we already know.’ In the case of a computer-generated text, the text may be compared to a human-written text, with expectations of syntactic and semantic sense informing the reader’s interpretation of computational output. Even if the reader is not told that the text is computer-generated, the sheer similarity of output to that which is human-written would be enough to prompt the same meaning-making processes applied to human-written texts. In these instances, scholarship associated with New Criticism, however endangered the approach may be, may offer support for reading practices. In New Critical thought, product is prioritised over process.

Process is especially difficult to discern when there is no clear identifiable author (Andersen, 2020). If, as per E. D. Hirsch (1967: 216), the ‘permanent meaning [of a text] is, and can be, nothing other than the author’s meaning,’ is the concept of authorial intention rendered obsolete for the computer-generated text? For an informational news item, the nature of

production may not be considered such an affront to understandings of authorship. GPT-2, however, is also capable of generating fiction. One abridged example follows (‘Better Language Models [Sample 5]’, 2019):

> System prompt (human-written): *Legolas and Gimli advanced on the orcs, raising their weapons with a harrowing war cry.*
>
> Model completion (machine-written, first try): The orcs’ response was a deafening onslaught of claws, claws, and claws; even Elrond was forced to retreat. ‘You are in good hands, dwarf,’ said Gimli, who had been among the first to charge at the orcs; it took only two words before their opponents were reduced to a blood-soaked quagmire, and the dwarf took his first kill of the night. The battle lasted for hours until two of the largest Orcs attempted to overwhelm Aragorn. When they finally stopped, they lay defeated and lifeless for miles and miles.
>
> ‘I take nothing,’ said Aragorn. ‘But I give my word, at my peril and mine, that I will never forget this day of horror. None of us will forget. Ever!’
>
> ‘I’ll never forget it!’ cried Gimli, who had been in the thick of the battle but hadn’t taken part in it. One of the wounded orcs he had carried off, he was the only one of the survivors who remained uninjured. ‘We’ll keep the memory of that day of evil, and the war with it, alive as long as we live, my friends!’

A more creative writing sample may prompt heightened senses of disillusion amongst readers. Creative endeavours such as fiction writing, after all, are supposedly safe in the event of a roboapocalypse. Yet the text above is readable, despite narrative inconsistency; one could conduct a traditional close reading (New Critique) of this text. In hermeneutics, though, context

must be considered alongside the text. Indeed,the context of this text's production – including the team of GPT-2's developers, the training corpus, and so forth – is especially relevant given that it differs so greatly from conventional publishing models. In a recent article asking "What do we lose when machines take the decisions?', Thomas Bolander asserts that GPT-2 shows no explicit signs of intention. Bolander (2019: 861) explains:

> The problem here is that the data we humans receive when learning language are far richer and contain far more dimensions than what has been fed into the language model. As children we learn language in the context of observing the objects and situations that the language refers to. The crucial importance of those additional dimensions is not necessarily obvious until one sees the consequence of training an algorithm without them.

In Boldaner's view, this lack of intention makes for a 'missing dimension' in textual output: a dimension that remains missing so long as intention appears absent. Once context like this is recognised, the reader can revisit the text to engage in reinterpretation informed by deeper and broader understanding. Hence, a hermeneutic circle is formed – around and around we go. Whether the reader identifies intention or not – or whether the reader chooses to prioritise intention during the interpretive process – is, as the history of hermeneutics has shown, altogether subjective.

A computer-generated text may be understood as having been produced by a genuine, albeit possibly synthetic, author. However, AI-driven authorship poses new challenges for hermeneutics with respect to how meaning originates, transmits, and develops. Are we to privilege the author or the reader, or adjust our understanding of a conventional author-reader relationship? The computational nature of AI-driven authorship also poses new opportunities

altogether: computational media (data and algorithms) with inherently traceable natures that contrast with the complex, liquid, ephemeral traits of human thought and its crystallisation in language. It is this traceable nature of the artificial that enables humans to precisely point at what pieces of training data, what steps in the algorithm, and what statistical relations and transformations are responsible for producing concrete characters, words, sentences, and even arguments and ideas in computer-generated texts. Leveraging these transformative traces might not just contribute to a more accurate genetic criticism, but might also be a powerful tool for understanding the processes of meaning-making undertaken by humans and, possibly, also machines. Word processing software has, as many genetic critics have lamented, obsolesced textual 'versions'. With some exceptions, the digitally-produced text leaves little evidence of its ontogenesis. For the computer-generated text, however, one can map, step by step, the journey of production. One can, at least potentially, identify the precise contributions of both human and computer.

At the same time, traceability faces issues in a digital age, namely to do with protection of intellectual property, or claims of ethical questionability. The non-release of GPT-2 – and eventually its gradual release, with OpenAI sharing the system in stages – inspired countless replication efforts. The best-known of these efforts is Nick Walton's freely-playable *AI Dungeon*, which transforms GPT-2's now open-source code into an interactive game wherein 'any thing [sic] you can express in language can be your action and the AI dungeon master will decide how the world responds to your actions' (Walton, 2019). As in early interactive computer games (e.g. *Zork*), players type text commands, to which the system responds. At times, the system's responses are humorous, providing content for full Twitter fan streams (Anonymous, 2019). Most of the time, however, the system responds in understandable and seemingly rational ways. As one journalist (Vincent, 2019) explains, 'the game doesn't really have a consistent sense of

who characters are or what they're doing. It just responds to the stimuli you feed it, line by line, rather than trying to follow long-term narratives or push concrete goals. In turn, this places the onus on the player. For the game to be any fun, you have to be imaginative and keep things going.' The reader is at least partially responsible for the output generated, and entirely responsible for that output's interpretation.

But who is the author of *AI Dungeon*'s output: the system; the system's user; Nick Walton; GPT-2's developers at OpenAI; some combination of the above; all of the above; none of the above? How is one to interpret the text that has been generated for individual users, in response to individualised input? Never mind what literary and/or cultural meaning and value such a text has. First, how do we trace where it came from, and who - or what - is responsible?

The issue of traceability is currently undergoing investigation in many fields. In the digital humanities, scholars working in 'tool criticism' aim to disentangle digital tools and algorithms, and clarify what software does to literary texts, how so, and why (Dorofeeva, 2014). Through such work, they hope to understand these tools' inner workings and embedded assumptions to explain how output was produced. This is done, for example, by identifying so-called 'data scopes' for digital history research: specific data transformations that can be used to track user interaction with research software (Hoekstra and Koolen, 2019). Instances of such tools being used include the aforementioned generation of texts by GPT-2, but also the generation of original music through variational autoencoders and deep learning (Roberts et al., 2018). In these cases, digital methods examine the statistical distributions and patterns of thousands of literary (or musical, respectively) examples; learn generalisations from these distributions and patterns; and then use these generalisations to generate yet unseen examples. The

fundamental question here is: how does this kind of process of artistic creation mimic or differ from that of human creators?

There might be an answer to this question in methods currently being investigated by computer scientists: methods addressing the reproducibility crisis, by which supplemental materials or 'reproducibility recipes' are integrated in generative methods to help others reproduce the same experimental results (Boettiger, 2015); the provision of better explanations for the AI system output (so-called eXplainable AI or XAI) (Sample, 2017); and the inclusion of provenance and workflow traces ('logs') that structurally model every step that a tool, algorithm, or model performs for content generation (Gil et al., 2013). All of these methods contribute to fine-grained historical pathways that explain the additions of and interactions with all human contributors to a system – invaluable information for genetic critiques of computer-generated texts. If we were to further include meaningful semantic relations to document all data interactions in generative artificial systems, we might be even closer to not only understanding how and why machines might be considered authors in themselves, but also how humans themselves establish conceptions of authorship (Kuhn et al., 2018). We would be closer to understanding the contexts of this new form of text. Through the application of hermeneutic theories, we could debunk computer-generated texts.

Training GPT-2 with Hermeneutics Texts

But to what extent do we need to fear Martin Eve's (2017: 46) post-anthropocentric world in which human writers have been replaced by language generation machines? So far, we have reflected on the impact of synthetically-generated texts in a range of quite specific human writing activities: fake news, short stories, or dungeon-like games. However, if Bolander's

postulates are true, and these language models lack fundamental dimensions that humans possess in their language learning processes, then there is really no post-anthropocentric intentionality to fear. To extend this argument, we suggest a tougher question for NLG. Can current language models like GPT-2 learn hermeneutic discourse from hermeneutics texts, and create novel – and even insightful – texts around that discourse?

To answer this question, we propose an experiment: re-train GPT-2 with articles and books cited throughout this paper, amongst others, and generate unconditional ('random') and conditional (in response to user-inputted prompts) samples on hermeneutics, and then analyse their discourse. This experiment is inspired by recent similar experiments using large bodies of scientific literature as training corpora for language models to see if these models could be used to assist scientific writing (Meroño-Peñuela et al., 2020). We collected 31 relevant papers on hermeneutics.[1] We then used the command line tool 'pdftotext' (Glyph & Cog, LLC.) to convert these papers into machine-readable plain text. Next, we used these texts, containing 541,568 words, to train the small model of GPT-2 over 48,700 training cycles until reaching a stable model. This process had two outcomes: numerous unconditional samples - one every 100 training cycles - and an interactive interface for conditional samples that can be publicly accessed by users to 'autocomplete' user-inputted prompts about hermeneutics (http://hermeneuticsgpt2.amp.lod.labs.vu.nl).

To better understand this program's generated output, let us recapitulate what GPT-2 actually is and how it works. GPT-2 is one example of a so-called transformer neural network architectures for deep learning. Transformers were made to address the problem of sequence transduction or neural machine translation. That is, given an input sequence ('The first law of

[1] Link to papers temporarily removed for peer review.

robotics states that a robot may not injure'), the system produces an output sequence of maximum probability ('a human being'). For this to work, some memory retention mechanism is necessary while the system is learning from the training set; most RNNs do this through message passing, chaining the output of one network to the next, but with decreasing quality of performance as sentences get longer and more complex. What sets GPT-2 apart from other language generators is its integration of a self-attention mechanism built on decoder blocks that makes it focus on only part of the information it is given. This means that GPT-2 can 'pay attention' to specific words, and build coherent 'arguments' around them.

So, what does our version of GPT-2 have to say about its own experiences of 'authorship' and the relationship between the author and the reader? Below are samples of output related to authorial intentionality:[2]

> (1) And often those things are not what I want to change, and the things that I actually say are often the things I see or feel about my work, and the things that I'm really passionate about or just about, in a very different way. It seems to me that that's a good sign that I'm not what a writer would like me to be, and I'm still not what someone who is passionate about my work would like me to be, or who is passionate about a certain book or a specific story.
>
> Some of these things might be different from what is actually happening to me, and I might have to change my thinking about those things or change my expectations about them.
>
> (2) No, what we are seeing is that for a text to be considered separate from its author, it will have to be seen in relation to the readers. This has been demonstrated by the way this practice of 'reading without thinking' has been used, if indeed any attempt be made to disentangle the different modes of thought. Consider this passage from Raymond Queneau's work Aeneid. The idea is that the reader can choose either to look for signs in the text or to 'go on' and 'hold on'. But the problem is that the reader can only look at the text on this 'reading' by itself, so he/she has to go on – thinking – to 'not look' for signposts in the text.

[2] See https://www.dropbox.com/s/je5cxq1arskhlgu/herm-gpt2.zip?dl=0 for the full dataset of unconditional samples.

> Of course, if it is found that the text cannot be separated from its author, this results in the reader being forced to rely on what he/she has read and so on to classify the textual succession; if the reader has no idea what he/she has been doing, then the author will be left in the dark and the text not being published. In short, while the idea of separating out an author's writing from his/her word or in his's word, does this mean that he/she can never be the real author of the text, and yet is he/she the real author of the text?
>
> (3) I would add, however, that the impulse which most aptly describes the works of literature is and has been primarily concerned with the business of telling a story, with a given plot, and with a promised outcome; it is particularly with this business that the business of telling a story as a social and historical business, that of turning a story into a political and social business, is struck.
>
> (4) We frequently talk about intention, but in very simple terms: the way you intend to go about writing tells us about your state of mind from a variety of sources, all of which are meaningful regardless of what the author may have meant by what he writes. I will not attempt that unified footing myself.

The themes of these samples are combinations and extensions of the themes of the papers used to train the program, and these samples were generated entirely in light of this curated scholarship. However, it would be an impossible feat to determine precisely which parts of a selected piece of output stemmed from which inputted paper. The bias of our curatorial process, and the program's subsequent bias reflecting the works upon which it had been trained, makes it so that issues related to authorial identity are not necessarily any less important, but are markedly more difficult to negotiate. Gender, race, socioeconomic status, and any other traits that may inform one's writing process - as well as a reader's interpretation of output - are diluted in the case of a computer-generated text. Even if the training set is biased and overrepresents one of these traits, the generated compositions are a coherent blend of the language model behind all of them. For this reason, much of the discussions in hermeneutics scholarship are called into question when applied to NLG programs like ours presented here. Our program's output avenues of thought that are not necessarily new, but are nevertheless original. The program itself encourages reflection not just on the content of the text, which could in many

cases be considered meaningful, but also the implicit biases that may be perpetuated through programming. These kinds of explicit discussions about the hermeneutics of computer-generated texts are vital as this genre evolves. There will never be any conclusive answers to questions of social value, reader interpretation, or authorial intention. But, in the words of our version of GPT-2: ‘Hermeneutics does not try to say that all is unknowable, but to insist that we should and can be certain, for even the things we say have an impact that is almost entirely theoretical, and if we are not certain we cannot express it.’

Conclusion

Hermeneutics underlies much of modern literary analysis, but more often than not goes unnamed. By explicitly recognising hermeneutics as part of textual interpretation, we are better suited for self-reflection upon personal processes of meaning-making. Hermeneutics scholarship offers a basis for a new theory of understanding in literary analysis: one that appreciates not just content, but form, and one that accommodates the vast array of publishing models now available in an increasingly digital context.

This article has been an appeal to re-introduce hermeneutics into literary study, particularly for its application to the burgeoning industry of computer-generated texts. It has explored the value of explicit application of hermeneutical tools like the hermeneutic circle for guiding consideration of both text and context. In his article about ‘Why Literature Professors Turned Against Authors’, Professor John Farrell (2019) observes that ‘it is impossible to get rid of authors entirely because the signs that constitute language are arbitrarily chosen and have no significance apart from their use.’ A literary study incorporating hermeneutics recognises this

impossibility, while at the same time appreciating the social contexts related to text production and to readers' individualised circumstances of meaning-making. Before rushing to engage in processes of interpreting new kinds of texts stemming from digital developments, we must critically self-reflect upon those processes. What cultural and literary assumptions are informing acts of meaning making? Are we able to reasonably compare these new kinds of texts with their more traditional counterparts, or do these new texts necessitate altered or altogether different evaluative criteria? We must consider assumptions about how literary studies are done, especially when digital technologies encourage reconsideration of conventional reading practices and conceptions of authorship. We must redirect attention to how meaning is made, overcoming reductionist views on stand-alone texts and embracing their increasingly available cultural contexts, interactions, and tools. Explicit discussions of hermeneutics, and engagement with past hermeneutics scholarship, facilitates calculated consideration of the future: a future circling around technological novelty and critical adaptability. Perhaps authorial intention has drifted from the uniqueness of God, to the multiplicity of humans, and now to collaborations with machines. To conclude this paper with the words of our version of GPT-2, generated in the visual style of a poem:

Much as I doubt nothing,
I say to the mind of every software
artist that he or she may be a part of
this complex but fruitful relationship.

References

Andersen J (2020) Understanding and interpreting algorithms: toward a hermeneutics of algorithms. *Media, Culture & Society*. OnlineFirst. Available at: https://doi.org/10.1177/0163443720919373.

Anonymous (2019) aidungeon_txt. *Twitter*. Available at: https://twitter.com/aidungeon_txt (accessed 9 April 2020).

Barthes R (1977) The Death of the Author. In: Heath S (trans) *Image, Music, Text*. London: Fontana Press, pp. 142-148.

Beer G (2014) The Reader as Author. *Authorship* 3(1). Available at: http://dx.doi.org/10.21825/aj.v3i1.1066 (accessed 7 April 2020).

Better Language Models and Their Implications [Sample 2] (2019) OpenAI. https://openai.com/blog/better-language-models/#sample2 (accessed 6 November 2019).

Better Language Models and Their Implications [Sample 5] (2019) OpenAI. https://openai.com/blog/better-language-models/#sample25 (accessed 6 November 2019).

Boettiger C (2015) An introduction to Docker for reproducible research, with examples from the R environment. In: *ACM SIGOPS Operating Systems Review, Special Issue on Repeatability and Sharing of Experimental Artifacts.* 49(1): 71-79. Available at: http://doi.org/10.1145/2723872.2723882.

Bolander T (2019) What do we lose when machines take the decisions? *Journal of Management and Governance*, 23: 849-867.

Dilthey W (1996) On Understanding and Hermeneutics: Student Lecture Notes (1867-68). In: Makkreel RA and Rodi F (eds), Makkreel RA (trans), *Hermeneutics and the Study of History (Selected Works: Volume IV*). Princeton: Princeton University Press, pp. 229-234.

Dorofeeva A (2014) *Towards Digital Humanities Tool Criticism.* Masters thesis, Leiden University, Netherlands.

Drucker J (2002) Theory as praxis: The poetics of electronic textuality. *Modernism/Modernity*, 9(4): 683-691.

Eagleton T (1996) *Literary Theory: An Introduction*, 2nd edition, Oxford: Blackwell, p. 66.

Ensslin A (2020) Hypertext Theory. In *The Oxford Research Encyclopedia of Literature*. Available at: https://doi.org/10.1093/acrefore/9780190201098.013.982 (accessed 8 April 2020).

Eve MP (2017) The Great Automatic Grammatizator: writing, labour, computers. *Critical Quarterly*, 59.3: 39-54.

Farrell J (2019) Why Literature Professors Turned Against Authors – Or Did They? *Los Angeles Review of Books*. Available at: https://lareviewofbooks.org/article/why-literature-professors-turned-against-authors-or-did-they (accessed 4 November 2019).

Fish S (1980) Literature in the Reader: Affective Stylistics. In: *Is There a Text in This Class?: The Authority of Interpretive Communities*. Cambridge, MA: Harvard University Press, pp. 21-67.

Fish S (1970) Literature in the Reader: Affective Stylistics. *New Literary History* 2(1): 123-162.

Fish S (2001) Yet Once More. In: Machor JL and Goldstein P (eds) *Reception Study: From Literary Theory to Cultural Studies*. New York: Routledge, pp. 29-38.

Foucault M (1998) What is an Author? In: Faubion JD (ed), Hurley R et al. (trans) *Aesthetics, Method, and Epistemology*. New York: The New Press, pp. 205-222.

Gadamer H-G (1988) On the Circle of Understanding. In: Connolly JM and Keutner T (eds and trans) *Hermeneutics Versus Science?: Three German Views*. Notre Dame, IL: University of Notre Dame Press, pp. 68-78.

Gil Y, Miles S, Belhajjame K, Deus H, Garijo D, Klyne G, Missier P, Soiland-Reyes S and Zednik S (2013) PROV model primer. *W3C Working Group Note*.

Grau M (2014) Putting Hermes Back into Hermeneutics. In: *Refiguring Theological Hermeneutics: Hermes, Tricker, Fool*. New York: Palgrave Macmillan, pp. 79-103.

Heidegger M (1996) *Being and Time: A Translation of* Sein und Zeit, trans. by Stambaugh J. Albany: State University of New York Press.

Henrickson L (2019) 'Towards a New Sociology of the Text: The Hermeneutics of Algorithmic Authorship' Empirical Studies. *Loughborough University Research Repository*. Available at: https://doi.org/10.17028/rd.lboro.c.4663709 (accessed 7 April 2020).

Hern A (2019) New AI fake text generator may be too dangerous to release, say creators. *The Guardian*. Available at: https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction (accessed 20 March 2019).

Hirsch Jr ED (1967) *Validity in Interpretation*. New Haven: Yale University Press.

Hoekstra R and Koolen M (2019) Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 52(2): 79-94.

Iser W (1989) *Prospecting: From Reader Response to Literary Anthropology*. Baltimore: The Johns Hopkins Press.

Jiang Z (2018) Is the 'Intention' There? On the Impact of Scientism on Hermeneutics. *European Review*, 26(2): 381-394

Kuhn T, Meroño-Peñuela A, Malic A, Poelen JH, Hurlbert AH, Ortiz EC, Furlong LI, Queralt-Rosinach N, Chichester C, Banda JM and Willighagen E (2018) Nanopublications: A growing resource of provenance-centric scientific linked data. In: *2018 IEEE 14th International Conference on e-Science (e-Science)*, pp. 83-92.

Martin N (2019) New AI Development So Advanced It's Too Dangerous To Release, Says Scientists. *Forbes*. Available at: https://www.forbes.com/sites/nicolemartin1/2019/02/19/new-ai-development-so-advanced-its-too-dangerous-to-release-says-scientists (accessed 20 March 2019).

McGann JJ (1991) *The Textual Condition*. Princeton: Princeton University Press.

Meroño-Peñuela A, Spagnuelo D, GPT-2 (2020) Can a Transformer Assist in Scientific Writing? Generating Semantic Web Paper Snippets with GPT-2. In: *17th Extended Semantic Web Conference* (ESWC 2020), posters & demos, Heraklion, Crete, 4 June 2020.

Moules NJ (2002) Hermeneutic Inquiry: Paying Heed to History and Hermes. *International Journal of Qualitative Methods*, 1(3): 1-21.

Ong WJ (1982; repr. 1997) *Orality and Literacy: The Technologizing of the Word*. London: Routledge.

Palmer RE (1969) *Hermeneutics: Interpretation Theory in Schleiermacher, Dilthey, Heidegger, and Gadamar*. Evanston: Northwestern University Press.

Pressman J (2018) The Novel in the Digital Age. In: Bulson E (ed) *The Cambridge Companion to the Novel*. Cambridge: Cambridge University Press.

Roberts A, Engel J, Raffel C, Hawthorne C and Eck D (2018) A hierarchical latent vector model for learning long-term structure in music. *arXiv preprint*. Available at https://arxiv.org/abs/1803.05428 (accessed 20 January 2020).

Sample I (2017) Computer says no: why making AIs fair, accountable and transparent is crucial. *The Guardian*, 5 Nov. Available at: https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial (accessed 29 November 2019).

Schleiermacher F (1998) *Hermeneutics and Criticism and Other* Writings, trans. and ed. by Bowie A. Cambridge: Cambridge University Press.

Shane J (2019) *You Look Like a Thing and I Love You: How Artificial Intelligence Works and Why It's Making the World a Weirder Place*. New York: Voracious.

Vincent J (2019) This AI text adventure game has pretty much infinite possibilities. *The Verge*. Available at:

https://www.theverge.com/tldr/2019/12/6/20998993/ai-dungeon-2-choose-your-own-adventure-game-text-nick-walton-gpt-machine-learning (accessed 10 April 2020).

Walton N (2019) *AI Dungeon*. Available at: https://aidungeon.io (accessed 9 April 2020).

Wimsatt Jr WK and Beardsley MC (1946) The Intentional Fallacy. *The Sewanee Review*, 54(3): 468-488.

Zimmerman J (2015) *Hermeneutics: A Very Short Introduction*. Oxford: Oxford University Press.