Humanist Studies & the Digital Age, 7.1 (2022)
ISSN: 2158-3846 (online)
http://journals.oregondigital.org/hsda/
DOI: 10.5399/uo/hsda/7.1.5

Unknown Future, Repeated Present: A Narrative-Centered Analysis of Long-Term AI Discourse

Micaela Simeone

Abstract: Recent narratives and debates surrounding long-term AI concerns–the prospect of artificial general intelligence in particular–are fraught with hidden assumptions, priorities, and values. This paper employs a humanistic, narrative-centered approach to analyze the works of two vocal, and opposing, thinkers in the field–Luciano Floridi and Nick Bostrom–to ask how the representational, descriptive differences in their works reveal the high stakes of narrative choices for how we form ideas about humanity, urgency, risk, harm, and possibility in relation to AI. This paper closely reads Floridi and Bostrom using different representational models and historical narratives from works in the environmental humanities, literary theory, bioethics, and the history of technology to uncover the imaginative terrain of recent long-term AI discourse and reveal the complexity and limitations of the messaging underlying the works of different authors.

As the digital sphere becomes increasingly populated by algorithms and machine-learning models, numerous scholars have shifted their focus beyond the present and towards the artificial intelligence (AI) of the future. Philosopher and Founding Director of the Oxford Future of Humanity Institute Nick Bostrom has been a prominent voice making the case for the pertinence of long-term AI concerns. Chiefly, Bostrom has discussed at length the prospect of machine superintelligence: AI that would supersede human-level general intelligence. We are, of course, still in the era of narrow AI: the human brain possesses many capabilities beyond those of the most powerful AI. In other words, all signs indicate that the potential for superintelligence lies dormant for now. However, Bostrom posits that artificial general intelligence–AI that can perform any intellectual task a human can–may arrive this century, and suggests that we might then see an intelligence explosion constituting a new shift in the knowledge substrate and resulting in superintelligence. Bostrom has been vocal about the theoretical consequences of such an AI, often arguing that a superintelligence could constitute an existential threat to humanity. Other academics conclude that long-term AI concerns are too indefinite and too far in the future to merit much discussion, arguing that our energies are better spent focused on short-term concerns stemming from the digital technologies that are already having profound impacts on our lives. Oxford Internet Institute professor Luciano Floridi has been one of the central voices in this camp, even calling discussions about a possible intelligence explosion irresponsibly distracting. Bostrom and Floridi's opposing voices constitute one region of long-term AI narratives, outside of many possible alternatives.

A close reading of the works of authors on both sides can reveal the complex, yet familiar representational possibilities shaping the conversation about AI risks. Since at least the 2014 publication of Bostrom's book *Superintelligence: Paths, Dangers, Strategies,* Bostrom and Floridi in particular have emerged as two of the more vocal thinkers in this conversation, and have often been positioned by sources as representatives of the opposing camps. For instance, A 2017 WIRED Italy article explains that on the subject of existential threat from AI, "experts are divided," and that "on the one hand, we find those who, like Bostrom, … warn about the existential risks of … AI, [while] on the other hand there are thinkers such as Luciano Floridi … who underline how this scenario is not absolutely plausible" (Signorelli, 2017).[1] In a 2014 review of Bostrom's *Superintelligence* as well as Floridi's *The Fourth Revolution* (2014), which is focused on the recent revolution in information and communication technologies (ICTs), philosopher John R. Searle spends most of the work strongly challenging both authors' theses, but in setting up his essay, he makes the point that "while Floridi celebrates the revolution, Bostrom is apocalyptic about the future" (Searle, 2014). It is clear that the works of Bostrom and Floridi have emerged as particularly useful points of entry into an analysis of both sides of the long-term AI discussion.

This paper examines the works of Bostrom and Floridi from a humanistic perspective with a literary-historical focus, uncovering the representational tensions within their clashing narratives of the future. Such a reading can help us consider the following: how do our representations of change, the artificial, and the limits of the human in the face of AI have long-standing historical origins stretching back at least as far as the Renaissance? How do the representational, descriptive differences across the works of Bostrom and Floridi reveal the high stakes of narrative choices for the way we conceptualize our technologies, their possibilities, and their risks? In a 2019 paper, authors Macnaghten et al. put forth a narrative approach to examining attitudes about emerging technologies, accomplishing the kind of rhetoric-focused analysis that this paper will attempt. Specifically, the authors argue for an approach that "examines how public attitudes are formed in relation to the interplay of wider cultural narratives about science and technology" (504). Similarly, this paper will examine how Bostrom and Floridi's attitudes may be formed in relation to the interplay of historical narratives about technology as well as narratives interrogated in other fields, particularly in literary theory and the environmental humanities. The narratives I identify within the Bostrom-Floridi dialogue have strong ties to seemingly unrelated aspects of culture, history, ethics, and science, and discerning these connections helps to uncover the imaginative terrain of recent long-term AI discourse.

Section I considers narratives around the technological possibility versus plausibility of superintelligence to show how Bostrom and Floridi exhibit fundamentally different positions regarding speculation. I argue that these narrative differences also reflect a rift between gradualism and catastrophism seen in other fields, as well as emphasize either human or technological risks. Section II considers how environmental humanist Rob Nixon's concept of "slow violence" offers a framework for unpacking the narrative problem presented by superintelligence; here, I argue that Bostrom leans towards a model of spectacular, sudden, and visible violence while Floridi centers slow, unspectacular, and often ignored violence. Section III looks at the moral layers of Bostrom and Floridi's narratives, showing how they suggest allegiances to either beneficent or non-maleficent approaches to technology; this difference is indicative of a broader, basic rift between optimistic and pessimistic outlooks. In Section IV, I argue that philosophy has been here before by recalling Renaissance conceptions around Technē to show how Floridi and Bostrom echo centuries-old narrative framings of technology and its relationship to humanity.

In sum**,** narrative possibilities for imagining the risks of our technologies include utopia and dystopia, optimism and pessimism, and catastrophism and gradualism; what's more, different narratives hoping to stake out what is urgent about technology can privilege modes of either spectacular violence or unseen violence as well as reflect varying moral priorities and faiths in humanity. Because of this, the narrative choices that we make when we choose to imagine certain technological futures implicitly determine which harms we address,

[1] Translation mine.

how soon we act, whose voices we listen to, and, ultimately, reflect how much we trust our own human judgement as the arbiters of our technological present and future. Additionally, if Floridi and Bostrom are two of the dominant voices shaping the long-term AI narrative, it is important to ascertain what expressive possibilities have been prioritized and which have been excluded or left unconsidered.

I: Probability, Gradualism, and Catastrophe

The academic conversation about superintelligence consists of clashing narratives about reality and about the urgency of our present. For Bostrom, one needs to look no further than the current state of the AI field for indications that longer-term concerns are worth considering: he emphasizes that while an intelligence explosion may not be imminent, because the field shows no signs of slowing down, AI safety measures and preventative strategizing are of utmost importance (322-3). Floridi's "now" on the other hand, is oriented around humanitarian problems that need solving, the current failures of the digital technologies that are already transforming our reality, and chiefly, our tendency to misuse them; Floridi posits that we need ethical technological solutions (rather than fears about AI) now more than ever to deal with these problems ("New Winter" 2). Underpinning these different conceptions of what is important seems to be two different narratives about the present: Bostrom's faith in the danger of AI is based on a fear of unchecked technology and AI progress, while Floridi's faith in the possibility of technological solutions is based on a fear of unchecked human failings. The narrative importance of this question of how much we are willing to center the human in conversations about technology will become clearer in later analyses focused on the status of the human in both authors' works–particularly, when discussing Floridi's narratives regarding human suffering and human ingenuity, and Bostrom's emphasis on humanity's flawed nature.

Different judgements regarding possibility also color the superintelligence conversation. In defending the intellectual project of making predictions about superintelligence in the face of undeniable uncertainty and fallibility, Bostrom writes, "refusing to offer probabilistic predictions would not make the epistemic problem go away; it would just hide it from view" (330). Floridi, on the other hand, tells us that superintelligence is utterly implausible, "a mere logical possibility," and therefore unworthy as a topic of serious discussion ("Singularitarians" 8). Underlying Bostrom's statement that the epistemic problem presented by superintelligence requires foresight is a model hinged on what fiction writer Amitav Ghosh has called "the centrality of the improbable" (Ghosh, 23); in this model, logical possibility is enough–the improbable possibility of superintelligence is enough to merit an imagination of its risks and therefore to merit reasoning and strategy. For Floridi, logical possibility is inadequate–the sudden and unimaginable, while possible, is not a priority, partly because it is the plausible (rather than the possible) that his model foregrounds. Already, we can start to see how narrative differences in the superintelligence conversation have wider-ranging impacts–a focus on possibility warrants speculation about superintelligence in the first place, while a focus on plausibility stops this future from being imagined.

Arguably, inherent in Bostrom's speculative project are extravagant images of AI. Though Bostrom tries to distance the image of superintelligence from associations with unrestrained, overdramatized robot takeover narratives–even dedicating numerous pages to the examination of gradual, unspectacular paths to superintelligence–simply considering the basic idea of an intelligence explosion requires a willingness to confront the unimaginable. Ghosh, in his 2016 book *The Great Derangement: Climate Change and the Unthinkable*, explores the form of the modern, realist novel alongside the history of gradualist geological narratives, writing that both came about through a "banishing of the improbable and the insertion of the everyday" (17). For

Ghosh, literary realism insists on telling stories that foreground the ordinary and relocate "the unheard-of toward the background" (17). Similarly, Ghosh explains, geologists have long navigated a divide between gradualist narratives that privilege a model where, as geologist Michael Rampino writes in *Cataclysms* (2017), "slow and steady geologic processes" create the major changes in geologic history, and catastrophist theories that center evidence related to "episodic catastrophic and sweeping changes" (14-15). A similar tension related to gradualism, catastrophism, and unthinkability characterize Bostrom and Floridi's different views on the intelligence explosion. Ghosh paraphrases the gradualist view as saying "nature does not make leaps" (20); Floridi makes a similar claim about AI when he criticizes the idea of an intelligence explosion by stressing that "a growth curve can easily be sigmoid" ("Singularitarians" 9). Unlike the sigmoid curve's rise and fall trajectory, catastrophe works at "unpredictable intervals and in most improbable ways" (Ghosh, 20). So does Bostrom's AI: Bostrom often warns about the possibility of unexpected flaws in seemingly foolproof designs, perverse instantiations, unexpected breakthroughs, and sudden radical transformations where, for instance, the world is transformed and "humanity [is] deposed from its position as apex cogitator over the course of an hour or two" (*Superintelligence*, 79). In sum, Ghosh helps us see that storytelling models related to the spectacular or the ordinary, and the gradual or the catastrophic inform how Bostrom and Floridi's respective messages follow well-established patterns for narratives about non-AI subjects.

Ultimately, examining how certain narrative themes such as gradualism and the unthinkable are embedded in Bostrom and Floridi's texts point to a kind of literary concern for both authors related to how messaging around long-term AI can take on new meaning, even to the point of distortion, when imbued with certain images and tropes. In a 2017 book review, Lisa Swanstrom adeptly paraphrases Ghosh, writing that in his assessment, "literary realism is especially ineffectual because it quashes speculation. Nothing is allowed to enter the realist narrative that will appear shocking, out of the ordinary, or out of sequence when woven into the primary narrative thread … [Ghosh] argues that such conventions fail us when we attempt to narrate our increasingly and shockingly unpredictable global climate" (402-3). For his part, Bostrom is similarly concerned by the limits of speculation, lamenting at one point that the "pioneers of artificial intelligence … did not contemplate the possibility of greater-than-human AI. It is though the speculation muscle had so exhausted itself in conceiving the radical possibility of machines reaching human intelligence that it could not grasp the corollary–that machines would subsequently become superintelligent" (*Superintelligence,* 5-6). While Bostrom indicates that open-minded speculation is important, he later stresses the need for attention to *how* exactly this speculation is mediated or accomplished; at one point, Bostrom tells us that a reasonable discussion about superintelligence is haunted by the anthropomorphized image of an evil superintelligent robot: "those inane Terminator pictures are taking a toll" (323).

Floridi, too, is concerned with how cultural images may distort thought in AI research, at one point critiquing Bostrom's project by positing that speculation is the unserious and distracted realm of science-fiction ("New Winter" 1). Ghosh describes how, in early geologist circles, catastrophist theories positing "upheavals of sudden and unimaginable violence" were denounced as seductive, sensationalist magazine headlines (22-3). Floridi has echoed this criticism, calling the futurologist perspective on AI "the curse of the airport bestseller" ("New Winter" 3). In both Floridi and Bostrom's arguments, then, a kind of literary concern focused on narrative choices and storytelling about technologies seems to underlie their perspectives. Floridi's faith in the urgency of the present and Bostrom's faith in the danger of AI share in common a kind of ethics for writing about the spectacular: we should be wary of the use of limitless imaginative variations as tools for conceptualizing the risks of technologies.

II: Representational Ethics and the Digital Realm

Narrative focus on the spectacular can perpetuate a kind of imbalance where such narratives wield more storytelling power and thus gain more traction than narratives of the slow, gradual, and unseen. In his 2011 book *Slow Violence and the Environmentalism of the Poor*, author and environmental humanist Rob Nixon urges environmental activists to pay attention to "slow violence" and explores how the reality of slow violence constitutes a representational, literary challenge. Nixon's concept of slow violence can do a lot to help unpack the narrative problem of long-term AI concerns and its stakes. First, it may be helpful to understand what slow violence is *not*: the opposite of slow violence is "violence [that] is customarily conceived as an event or action that is immediate in time, explosive and spectacular in space, and as erupting into instant sensational visibility" (Nixon, 2). This spectacular, explosive, and instantly visible violence is the landscape of Bostrom's intelligence explosion. Bostrom does acknowledge that existential catastrophe need not be the default outcome of the creation of a superintelligence. However, his work emphasizes how the first superintelligence may have the advantages necessary to shape the future of life on Earth, could easily have non-anthropomorphic goals, and would likely have instrumental reasons to pursue open-ended resource acquisition; in light of these considerations, Bostrom tells us, "we can see that the outcome could easily be one in which humanity quickly becomes extinct" (141). Though he does distinguish three classes of transition scenarios–slow, moderate, and fast (77)–Bostrom urges us not to dismiss the possibility of a sudden radical transformation in artificial general intelligence as fanciful, telling us that "if and when a takeoff occurs, it will likely be explosive" (79). This explosive transformation, according to Bostrom, would result in incredible geopolitical, social, and economic upheaval, and would possibly result in a singleton, a single machine superintelligence project which has attained a decisive strategic advantage, "a level of technological and other advantages sufficient to enable it to achieve complete world domination" (96). Bostrom's vision of the singleton, then, would achieve the kind of spectacular violence Nixon describes; the singleton comes into violent being explosively and spectacularly, making itself at once visible and permanent.

Upon attaining its strategic advantage, Bostrom's singleton could, if it were motivated to do so, eliminate the human species in a single "strike," or it could immediately embark on a massive global construction project suitable to the realization of its own goals but resulting in humanity's demise due to widespread habitat destruction (Bostrom, 118). At almost every opportunity, Bostrom's singleton scenario resists any gradualizing, de-dramatizing effects. At one point, Bostrom posits that even those strategies which may have previously worked with a seed AI (a model intended to become superintelligent) could suddenly backfire in what he calls "the treacherous turn" where a previously cooperative AI, once it becomes stronger, strikes and forms a singleton without warning or provocation (144). The violence of the superintelligence of Bostrom's imaginative project is both spectacular and instantaneous; Nixon can help us see how these qualities give the singleton scenario a certain narrative power. First, however, it will be helpful to explore the particular technological "slow violence" which this spectacular violence model renders invisible, and which I argue Floridi hopes to bring into the forefront.

At the core of Nixon's slow violence concept is a compelling argument for including marginalized voices into existing narratives that privilege the spectacular. Environmental discourse, like AI discourse, would benefit from those perspectives that are often dismissed, such as perspectives from the Global South. Nixon argues that we need to understand a different kind of violence from that sensational violence we are used to: "a violence that is neither spectacular nor instantaneous, but rather incremental and accretive, its calamitous repercussions playing out across a range of temporal scales" (2). This violence is what Nixon calls "slow violence": a "violence that occurs gradually and out of sight, a violence of delayed destruction that is dispersed across time and space, an attritional violence that is typically not viewed as violence at all" (2). As we have seen, Floridi has been one of the most vocal critics of the conversation about machine superintelligence, and about the singleton "takeover" scenario in particular. Now, I want to argue that Floridi's criticism of superintelligence speculation takes on a new narrative dimension: we can read Floridi's call to attention regarding present-day technology problems as a call to recognize the slow violence that is afflicted by our digital artifacts.

In discussing past U.S and European international environmental schemes, Nixon locates the harms caused by these schemes within the sphere of slow violence. He explains that with slow violence's quiet, long-term casualties and invisible, unspectacular harms, the victims of this kind of violence are often silently displaced; on the international stage, for example, richer nations come to view the harms inflicted on citizens of poorer nations as "out-of-sight" and "remote from … activists' terrain of concern" (2). Floridi leverages a very similar critique against those he calls the "singularitarians"–those thinkers who, like Bostrom, are concerned by the prospect of an intelligence explosion resulting in superintelligence ("Singularitarians" 8). In a 2020 article, Floridi argues that thinkers who frame AI as humanity's greatest existential risk speak "as if most of humanity did not live in misery and suffering. As if wars, famine, pollution, global warming, social injustice, and fundamentalism were science fiction, or just negligible nuisances, unworthy of their considerations" ("New Winter" 1-2). Here, Floridi suggests that the profound risk of the spectacular violence model of superintelligence is that it erases the everyday but extremely urgent and *human* problems. Floridi mirrors Nixon in these criticisms. For Floridi, the focus on superintelligence positions the current-day sufferings of humanity as out-of-sight, thereby discounting the long-term, nebulous, and often silent violence that accompanies these sufferings in favor of the spectacular narrative of the AI singleton.

Floridi wants to redirect attention to the already-pertinent challenges of our digital technologies; he sums up this critique when he claims, "the futurologists find these questions boring" ("New Winter" 3). Floridi's criticism that the singleton narrative pushes aside the "boring" concerns of present-day technologies suggests that the digital realm lends itself to a Nixon-esque slow violence. For instance, we might ask, *how does bias in AI operate as a form of slow violence?* AI research lab OpenAI's third generation Generative Pre-Trained Transformer (GPT-3), a machine-learning model that uses its deep-learning experience to generate human-like text, is a good example of how reading technology harms in terms of slow violence may be more beneficial than catastrophizing. GPT-3 learns language by studying the statistical patterns in a dataset of roughly one trillion words collected from the web (Brown et. al, 8). However, researchers have found that biases present in GPT-3's vast training corpus of Internet text has led the AI to internalize these patterns and generate stereotyped and prejudiced content as a result (Brown et. al, 36). Nixon describes slow violence as leaving behind "hushed havoc" which "eludes … tidy closure … [or] containment" (6). Similarly, GPT-3's internalization of the Internet's nebulous web of social, political, and sentimental patterns happens quietly–within the opaque layers of GPT-3's neural network–and leaves behind harms that elude any easy fix. These harms are felt most by those at the margins, both in terms of minimal representation in the model and in terms of minimal control over AI systems; these are the same populations that are impacted disproportionately by the other kinds of slow violence which Nixon describes.

GPT-3 is not yet publicly available; however, we do not need to look beyond our already ubiquitous digital artifacts to locate the slow violence of AI bias. A 2013 paper by linguistics scholars Paul Baker and Amanda Potts explores the bias of Google Autocomplete algorithms which internalize queries to do their job, but in doing so often reflect "-ism" statements and biases when generating suggestions based on common searches

(187). Nixon further describes the slow-moving and long-in-the-making disasters of slow violence as disasters "that are anonymous and star nobody" (3); the casualties of these disasters typically pass untallied and unremembered. Such discounting in turn makes it far more difficult to "secure legal measures for prevention, restitution, and re-dress" (9). A similar discounting is accomplished by the kind of AI bias Baker and Potts describe. The silent embeddedness of algorithmic bias means that the targets of digital prejudice are almost never named and made visible. The scalability of such algorithms and their ubiquitous but invisible operation within the architectures of the Internet mean that prevention and restitution for their harms is a complex and elusive project (Noble).

By analyzing the digital as a realm of slow violence, we are better able to see how Floridi's call to action is one that asks us to reckon with this slow violence, the hidden but already urgent human struggles that are shaped by our own misuse of digital technologies ("Singularitarians" 10). When juxtaposed with the slow violence of the present-day digital sphere, the narrative power of Bostrom's singleton scenario becomes all the clearer. A superintelligence takeover scenario constitutes a representational success in what Nixon describes as an age "when the media venerate the spectacular" (3). Like those environmental spectacles Nixon describes– "burning towers, … avalanches, volcanoes, and tsunamis" (3)–the image of the superintelligence singleton also has a "visceral, eye-catching, and page-turning power that tales of slow violence, unfolding over years, decades, even centuries, cannot match" (3). For Floridi, it seems, our project should be to infuse the amorphous calamities of global, long-standing human sufferings and digital failings with the dramatic urgency of the superintelligence narrative, but with an eye towards solutions and not organized around fears. For Floridi, "AI must be treated as a normal technology, neither as a miracle nor as a plague, and as one of the many solutions that human ingenuity has managed to devise" ("New Winter" 2). Ultimately, in Floridi's view, listening to the spectacular superintelligence narrative stops us from reckoning with the current, non-spectacular harms inflicted by humans and taking place in a present that is already being reshaped by our technologies in profound ways.

In a 2016 paper, "Invasive Narratives and the Inverse of Slow Violence," authors Lidström et al. consider Nixon's argument that "a lack of 'arresting stories, images, and symbols' reduces the visibility of gradual problems … in cultural imaginations and on political agendas" (1). This is essentially the same concern that Floridi has about the superintelligence narrative—for him, the pull to imagine a spectacular intelligence explosion is arresting and intoxicating, and diverts energies from focusing on solutions. Notably, Lidström et al. call the problem of difference between arresting stories and slow violence–which does not capture attention in the same way–a "representational imbalance" (1). For Floridi, addressing this problem means framing AI as a normal technology and as a solution. The idea of representational imbalance provides useful vocabulary for emphasizing how action and philosophy around AI can be shaped by rhetorical choices. Lidström et al. add on to Nixon's work, arguing that the tools leveraged by the imbalance itself can be part of how slow violence can become foregrounded; the authors advocate an approach that "unpack[s] the ways that complicated and multifaceted … phenomena can be reduced to fast, simple, evocative, *invasive* narratives that percolate through science, legislation, policy, and civic action" (1). Floridi's criteria that AI be treated "neither as a miracle nor as a plague" seems to preclude the possibility for an evocative, invasive narrative, but it is a possibility worth considering. Regardless, it is clear that the Bostrom-Floridi, long-term/short-term AI rift is partly also one between narratives and priorities related to different kinds of violence and harm.

These considerations require us to ask where spectacular narratives about AI may also result in a beneficial framing of AI concerns. This question brings us back to Ghosh, particularly, to his assertion that we are, in the context of climate change, living in a time of distinctive events that resists the imposition of a false and immobilizing narrative of regularity. Ghosh posits that "we are now in an era that will be defined precisely by events that appear, by our current standards of normalcy, highly improbable" (24). Elsewhere, he writes that "the Anthropocene has already disrupted many assumptions that were founded on the relative climatic stability of the Holocene" (21). Similarly, Bostrom and other superintelligence thinkers ask us to consider how previously extraordinary technological conditions we now take to be ordinary suggest that we would benefit from at least considering worst-case AI scenarios in the absence of certainty that they will never come into fruition. Stuart Russell, Berkeley professor and AI researcher, has argued in favor of the superintelligence conversation, pointing out in 2015 that scientific breakthroughs can often surprise researchers in the field themselves: "some have argued that there is no conceivable risk to humanity [from AI] for centuries to come, perhaps forgetting that the interval of time between Rutherford's confident assertion that atomic energy would never be feasibly extracted and Szilárd's invention of the neutron-induced nuclear chain reaction was less than twenty-four hours" (Russell, 2015). Similarly, Bostrom recalls unexpected leaps made by AI; at one point, he writes that "it was once supposed ... that in order for a computer to play chess at grandmaster level, it would have to be endowed with a high degree of general intelligence (...). It turned out to be possible to build a perfectly fine chess engine around a special-purpose algorithm" (14). Bostrom's underlying point here is that some of our earlier assumptions about human intelligence have been proven wrong by AI in the past, so future dramatic shifts in AI's relationship to human intelligence cannot be entirely discounted. Bostrom is also careful to note that AI is already outperforming human intelligence in certain domains, and that if these achievements do not seem impressive to us it is only because "our standards for what is impressive keep adapting to the advances being made" (14). The digital realm, in this view, is not characterized by slow violence but is instead increasingly marked by distinctiveness and previously discounted, marvelous achievements. In this model, the project of assessing implications of future technologies *before* they become feasible requires a willingness (justified by history) to engage with the *possibility* of the spectacular, the sudden, and the immediate, however improbable.

There is a complication to this reading, however: In his November 2015 Afterword in a republished edition of *Superintelligence*, Bostrom is careful to note that his argument is not that an intelligence explosion is imminent, but that it is pertinent nevertheless to consider the improbability of a recurrent AI winter as severe as those the field has experienced in the past (322) due to funding and reputability flowing into the field which show no signs of slowing. So, while Bostrom does seem to align himself with a representational ethics of the spectacular, this moment from his Afterword suggests that he really feels that the main loss of the Floridi argument (*we ought to focus our energies entirely on the slow violence of the here and now*) is that we do not fully reckon with the reality of AI's incremental but notable progress. As discussed, it seems that Floridi would see the digital as engaged in a kind of slow violence before our eyes, and Bostrom's entire imaginative project regarding superintelligence is predicated on a conviction that AI is incrementally becoming formidable. So, in addition to the shared dislike of fully imaginative, science fiction relationships to AI concerns, there is another strange similarity between Bostrom and Floridi's narratives: both authors do suggest that our current digital landscape is making gradual progress worth paying close attention to.

III: Moral Narratives

Thus far, I have only briefly touched on how Floridi's argument navigates our focus away from fears of the dangers of AI and towards the urgency of technological solutions. This consideration can lead us to see how different moral priorities are embedded in Bostrom and Floridi's different representational choices. Bostrom's narrative reveals a prioritization of non-maleficence, while Floridi's privileges beneficence.

Both Bostrom and Floridi are concerned with the ethical and moral implications of digital technologies. However, underlying both authors' thoughts about how we might ensure that digital technology serves the interests of humanity and promotes the common good is a distinction between the negative principle ("do no harm") and the positive principle ("do good") in AI development–principles which, notably, are widely applied in bioethics (McCormick). Put another way, approaches toward developing principled AI can be either about ensuring that those systems are beneficent or ensuring they are non-maleficent. Though Bostrom does present a "common good principle" which says that "superintelligence should be developed only for the benefit of all humanity and in the service of widely shared ethical ideals" (312), his overarching sentiment is that we ought to be very careful with AI development in order to avoid the wide-ranging harm possible with general machine intelligence. Floridi takes the opposite approach when he accuses those concerned with superintelligence of misleading public opinion to be fearful of AI progress rather than knowledgeable about the potential and much-needed solutions AI could bring about. Echoing the beneficence principle, Floridi writes, "we need all the good technology that we can design, develop, and deploy to cope with these challenges, and all human intelligence we can exercise to put this technology in the service of a better future" ("New Winter" 2). In his afterword, Bostrom echoes the non-maleficence principle when he writes, "whereas we might get by with a vague sense that there are (astronomically) great things to hope for if the machine intelligence transition goes well, it seems more urgent that we develop a precise detailed understanding of what specific things could go wrong—so that we can make sure to avoid them" (*Superintelligence* 324).

There are risks with both models. With the beneficence approach, one potentially risks sidelining preventative AI safety mechanisms in the pursuit of more pressing technological solutions, resulting in what Bostrom calls an "AI safety winter" (323). Arguably, the beneficence model's dedication to socially good AI requires a commitment to comprehensively answering the thorny question, "whose common good?" These considerations aside, the beneficence approach seems to center humanitarian motives in ways the non-maleficence approach may overlook. Bostrom's harm-centric view of long-term AI may lead to somewhat impersonal calculations of how best to avoid specific risks or outcomes. In a final chapter on strategic considerations, Bostrom calls for a focus on elastic problems, "problems that can be solved much faster, or solved to a much greater extent, given one extra unit of effort" (316). He explains that while "achieving world peace … would be highly desirable, … considering the numerous efforts already targeting that problem, and the formidable obstacles arrayed against a quick solution, it seems unlikely that the contributions of a few extra individuals would make a large difference" (316). Floridi, it seems, would take the opposite view, to argue instead that the development of socially good AI in the service of the common good should be where we focus our energies. To get to Bostrom's spectacular singleton scenario, we need to embark on an imaginative exercise that centers risk and attempts to formulate all possible worst-case futures. To get to Floridi's vision of socially good AI, our imaginative exercise instead centers optimism about technology solutions in the first place and the beneficent, useful possibilities of digital technologies and AI.

IV: Comparable Renaissance Narratives on the Works of Technē

This simple but fundamental difference in the imaginative exercises of Bostrom and Floridi–pessimism versus optimism–has origins stretching at least as far back as the European Renaissance (1450-1700), predating the Industrial Revolution and any recognizable relationship to the technologies of today. Renaissance thinkers navigated the increasing presence of mechanism in their lives with similar representational and narrative differences to those within the long-term AI discussion. The similarities can reveal what has always been at stake in the work of creating narratives and imagining possibilities for technology and its place in the world.

The tension between "art" and nature is a defining theme within the works of Renaissance philosophers, artists, and authors. The term "art" was used to describe the man-made world of devices, instruments, and other human interventions in nature. Historian Jonathan Sawday, in his 2007 book *Engines of the Imagination: Renaissance Culture and the Rise of the Machine*, writes about the art/nature distinction which shaped attitudes towards the artificial world of Renaissance Europe. Sawday tells us that over the course of the fifteenth, sixteenth, and seventeenth centuries, as increasingly complex mechanical devices and other human arts came into being, two conflicting attitudes–one optimistic and one pessimistic–arose from the "Aristotelian distinction between the natural world and the artificial world of Technē" (Sawday, 3). Some thinkers saw the works of Technē as the expression of human ingenuity working alongside nature to bring about a better and more useful world. At the same time, others perceived the arts with more skepticism, seeing many productions of art as vain attempts to shape nature. The optimists viewed the works of art as man's ultimate accomplishment, while the pessimists viewed them as markers of man's fallen, labor-bound state (Sawday, 3-4).

Centuries later, similar drives to wrangle the productions of technology into either utopian or dystopian narratives persist. While the opinions of Bostrom and Floridi are altogether more nuanced than optimism and pessimism, there are aspects of their positions which reveal similar fundamental hopes or despondencies about man's relationship to technology. As we have seen, Floridi in part optimistically aligns human ingenuity with technology, writing that AI must be treated as "one of the many solutions that human ingenuity has managed to devise" ("New Winter" 2). Sawday writes about the optimistic view of man-made devices during the Renaissance that these thinkers felt that such devices promised a partial theological restitution–the original disaster of the Fall of humanity in Eden could be alleviated through Technē "working in the service of humankind and understood as a product of human ingenuity" (3). Though Floridi's narrative does not include this religious dimension, his view of AI seems to be bolstered by a similar faith in the controllability of technology–AI models are extensions of *human* vision and *human* craft, and thus can be *made* to work on our behalf and reflect our values in order to rectify or repair some of the damage humans have caused. Floridi also writes that our planetary problems, including global warming, social injustice, and migration, would benefit from AI solutions because AI can give us "increasingly smarter ways of processing immense quantities of data, sustainably and efficiently" (2). This statement reflects another optimism about AI: the smarter the AI, the more potential for good. In other words, Floridi expresses a faith in the power of AI's unique form of intelligence to solve humanity's problems. In general, Floridi's faith seems to lie in AI's potential ability to complement human ingenuity, intelligence, and reason rather than inherently reflect our faults.

Bostrom's writings communicate a different narrative–one that aligns itself instead with Sawday's description of the pessimistic view of Renaissance devices. Sawday tells us that the pessimists saw Technē as conflicting with nature and even God, and felt that "despite their usefulness, machines were products of fallible human reason, and as such they were always to be considered as tainted in some way" (4). Similarly, Bostrom also often pessimistically aligns technology with human error and inadequacy. In describing incentive-based methods for controlling the motivations of a seed AI, Bostrom notes that one problem with this path is that the AI may not trust the human operator to deliver the promised rewards because, after all, "the track record of human reliability is something other than a straight line of perfection" (163). This is one example of Bostrom taking care to account for all the possible ways human error and fault could make their way into the technological architecture of a superintelligence. In a later section, Bostrom writes that "human nature … is flawed and all too often reveals a proclivity to evil which would be intolerable in any system poised to attain a decisive strategic advantage" (232). Interestingly, to address this issue, Bostrom ends up recommending the design of an unabashedly artificial system rather than a more neuromorphic one (232). So, not only does Bostrom echo the Renaissance pessimist view that the works of Technē are at risk of reflecting human defect, but he also takes a step farther and expresses a willingness to decenter the human as the intellectual center of the universe, advocating a totally artificial alternative. The confrontation of the possibility that machines may rival humanity's natural capabilities also has origins at least as far back as the European Renaissance. In a 2007 essay, scholar of Shakespeare and Renaissance humanism Scott Maisano writes that Renaissance automata–mechanical, self-moving, mechanical objects that populated the period's royal gardens and other estates, and which were fascinations and marvels to numerous key philosophical figures of the time, including Francis Bacon and Descartes–were "fantasy figures of consistency, competency, and self-control" which possessed levels of "endurance and efficiency … [that] set the (impossible) standard for human conduct" (75).

Bostrom's willingness to de-anthropomorphize perhaps gets at the core of his narrative project as compared to Floridi's. For Bostrom, the future is artificially intelligent–less-than-genius humans will need to recognize our shortcomings in the face of increasingly powerful AI and contribute this awareness to our AI design processes if we are to ensure that AI remains on our side. Conversely, for Floridi, the future is human–digital technologies can work on our behalf and they can be subject to misuse but either way, the momentum behind any change will be human. Renaissance thinkers were confronting many philosophical upheavals as the onset of the Scientific Revolution while new devices began to shift humans' relationship to nature. Today, we are confronting a digital sphere that is reshaping our relations to one another, our political lives, our conceptions about what it means to be human, and more. Renaissance philosophers found out that grappling with these issues involves a reckoning with either our utopian hopes or our dystopian fears. Whether we side with Bostrom, Floridi, or land somewhere in the middle, the narrative choices we make in our imaginings of AI will shape considerations about humanity's place in a technology-driven world.

Conclusion

A narrative-centered analysis of long-term AI discourse and its historical precedents reveals how certain narratives and approaches are privileged over others, such as Bostrom's spectacular singleton narrative or Renaissance Europe's emphasis on human ingenuity in the mechanical arts. In addition, reading Floridi alongside Nixon shows how certain narratives can erase many experiences and perspectives and why AI discourse, like environmental discourse, ought to account for marginalized narratives. However, it is important to acknowledge that both Floridi and Nixon (as well as Bostrom and others cited in this paper) represent historically privileged perspectives–white, male, European, American. In conclusion, it is worth noting that a reading like this one which seeks to uncover the narrative challenges and possibilities surrounding the topic of long-term AI risks moves us to ask who is speaking. In other words, how might other voices offer alternative narrative possibilities to the white, patriarchal, European perspectives represented by thinkers like Floridi and Bostrom? Floridi does continually remind us of the other voices needed in conversations about technology's impacts–particularly, voices from marginalized and disenfranchised groups–and Nixon's *Slow Violence* also makes clear the need for a conception of violence that makes visible the often ignored injustices impacting, particularly, communities in the Global South. It has been this paper's purpose to show that a narrative-focused analysis of discourse about long-term AI risk ultimately leaves us with an understanding of how representational choices alter our conceptual relationships to our technologies in real and impactful ways. Once we acknowledge this, it becomes clear that representational models from non-white, feminist, queer, Indigenous, and Global South perspectives, to name a few, could point to alternative futures yet to be imagined or recognized by those thinkers who have been driving our technology narratives thus far. One example of an alternative model is presented by scholar of law and AI ethics Suvradip Maitra in a 2020 paper that considers the value of Indigenous epistemologies to a conversation about increasingly general AI. In particular, Maitra highlights how numerous Indigenous perspectives offer "well-developed epistemologies adept at accounting for the non-human, a task that defies Western anthropocentrism," and that would thus potentially provide useful steps toward finding new ways to frame our relationship with AI as AI models become more powerful and their capabilities challenge our conceptions of human and machine (320). These reflections get to the core of what is at stake in the narrative choices made when imagining AI futures and contemplating our digital present: diverse outlooks on the possibilities of our technologies and varying faiths about humanity's role in realizing them.

Works Cited

Baker and Potts. "'Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms." *Critical Discourse Studies*, vol. 10, no. 2, 2013, pp. 187-204. https://dx.doi.org/10.1080/17405904.2012.744320

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*, 2014, Oxford University Press, Rpt. 2017. Print.

Brown et al. "Language models are few-shot learners." arXiv preprint, 2014, arXiv:2005.14165.

Floridi, Luciano. "AI and Its New Winter: from Myths to Realities." *Philosophy & Technology*, vol. 33, 2020, pp. 1-3, https://doi.org/10.1007/s13347-020-00396-6.

Floridi, Luciano. "Singularitarians, AItheists, and Why the Problem with Artificial Intelligence is H.A.L (Humanity At Large), not HAL." *APA Newsletter on Philosophy and Computers*, vol. 14, no. 2, 2015, pp. 8-10. Print.

Ghosh, Amitav. *The Great Derangement: Climate Change and the Unthinkable*, University of Chicago Press, 2016. Print.

Lidström et al. "Invasive Narratives and the Inverse of Slow Violence: Alien Species in Science and Society," *Environmental Humanities*, vol. 7, no. 1, May 2016, pp. 1–40, https://doi.org/10.1215/22011919-3616317

Macnaghten, Phil, et al., "Understanding Public Responses to Emerging Technologies: A Narrative Approach", *Journal of Environmental Policy & Planning*, vol. 21, no. 5, 2019, pp. 504-518, https://doi.org/10.1080/1523908X.2015.1053110

Maisano, Scott. "Infinite Gesture: Automata and the Emotions in Descartes and Shakespeare," *Genesis Redux: Essays in the History and Philosophy of Artificial Life*, edited by Jessica Riskin, University of Chicago Press, 2007. Print.

Maitra, Suvradip. "Artificial Intelligence and Indigenous Perspectives: Protecting and Empowering Intelligent Human Beings," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20),* Association for Computing Machinery, New York, NY, 320-326, https://doi.org/10.1145/3375627.3375845

McCormick, Thomas R., "Principles of Bioethics," *University of Washington Medicine Department of Bioethics and Humanities*, University of Washington, Accessed 25 Sep 2021, https://depts.washington.edu/bhdept/ethics-medicine/bioethics-topics/articles/principles-bioethics

Nixon, Rob. *Slow Violence and the Environmentalism of the Poor*, Harvard University Press, 2011. *ProQuest*. Print.

Noble, Safiya. *Algorithms of Oppression*, NYU Press, 2018.

Rampino, Michael, R., *Cataclysms: A New Geology for the Twenty-First Century,* Columbia University Press, 2017, https://doi.org/10.7312/ramp17780.

Russell, Stuart. "2015: What do you think about machines that think?" *Edge*, 2015, www.edge.org/response-detail/26157.

Sawday, Jonathan. *Engines of the Imagination: Renaissance culture and the rise of the machine*, Routledge, 2007. Print.

Searle, John R., "What Your Computer Can't Know," *The New York Review*, 9 October 2014, https://nybooks.com/articles/2014/10/09/what-your-computer-cant-know/

Signorelli, Andrea, D., "Tre leggi per regolare l'intelligenza artificiale," Translation mine, WIRED Italia, 13 Oct 2017, https://www.wired.it/gadget/computer/2017/10/13/tre-leggi-regolare-intelligenza-artificiale/

Swanstrom, Lisa, "Serious Literature? Science Fiction at the MLA," *Science Fiction Studies*, vol. 44, no. 2, SF-TH Inc, 2017, pp. 402-404, https://doi.org/10.5621/sciefictstud.44.2.0402.