Machine Learning in Contemporary Science Fiction

Jo Walton

> "To suggest that we democratize AI to reduce asymmetries of power is a little like arguing for democratizing weapons manufacturing in the service of peace. As Audre Lorde reminds us, the master's tools will never dismantle the master's house." –Kate Crawford, *Atlas of AI*
>
> "Why am I so confident?" –Kai-Fu Lee, *AI 2041*

Suppose There are Massacres

Suppose there are massacres each day near where you live. Suppose you stumble on a genre of storytelling that asks you to empathize with the *weapons* used by the murderers. Confused by this strange satire, you ask the storytellers, 'What's the point of pretending these weapons have inner lives?' They readily explain, it is mostly just for fun. However, there are serious lessons to be learned. For example, what if 'we' — and by 'we' they mean *both* the people wielding the weapons, *and* the people getting injured and killed by them — what if *we* one day *lost control* of these weapons? Also, in these stories, the anthropomorphic weapons often endure persecution and struggle to be recognized as living beings with moral worth… just like, in real life, the people who are being massacred!

Disturbed by this, you visit a nearby university campus, hoping to find some lucid and erudite condemnations, and maybe even an explanation for the bizarre popularity of these stories. That's not what you find. Some scholars are obsessed with the idea that stories about living weapons might somehow influence the development of real weapons, so much so that they seem to have lost sight of the larger picture. Other scholars are concerned that these sensationalizing accounts of the living weapons fail to convey the many positive impacts that similar devices can make. For example, a knife has uses in cooking, in arts and crafts, in pottery, carving away excess clay or inscribing intricate patterns. In snowy peaks, a bomb can trigger a controlled avalanche, keeping the path safe for travelers. In carpentry or in surgery, a saw has several uses. Even the microwave in your kitchen, the GPS in your phone, and diagnostic technologies in your local hospital have origin stories in military research. These are only a few peaceable uses of weapons so far, the scholars point out, so imagine what more the future may hold. Eventually you do actually find some more critical perspectives. But you are shocked you had to search so hard for them.

Science Fiction and Cognition

The small preamble above is science fiction about science fiction. Just as science fiction often aims to show various aspects of society in a fresh light, this vignette aims to show *science fiction about AI* in a fresh light. The reason for talking about weapons is not just that AI is directly used in warfare and genocide, although of course that's part of it. But the main rationale is that the AI industry is implicated in a system of slow violence, one which perpetuates disparities in economic inequality, and associated disparities in safety, freedom, and well-being. It is part of a system whose demand for rare minerals threatens biodiversity and geopolitical stability, and whose hunger for energy contributes to the wildfires, famines, deadly heatwaves, storms, and other natural disasters of climate change. These are not the only facts about AI, but they are surely some of the more striking facts. One might reasonably expect them to loom large, in some form or other, in science fiction about AI. However, in general, they don't.

This vignette is written to challenge a more optimistic account of science fiction about AI, which might go as follows: science fiction offers spaces to examine the social and ethical ramifications of emerging AI. As a hybrid and multidisciplinary discourse, science fiction can enliven and energize AI for a range of audiences, drawing more diverse expertise and lived experience into debates about AI. In this way, it may even steer the course of AI technology: as Chen Qiufan writes, speculative storytelling "has the capacity to serve as a warning" but also "a unique ability to transcend time-space limitations, connect technology and humanities, blur the boundary between fiction and reality, and spark empathy and deep thinking within its reader" (Chen 2021, xx). Anticipatory framings formed within science fiction are also flexible and can be adapted to communicate about and to comprehend emerging AI trends. Of course, science fiction is not without its dangers; for example, apocalyptic AI narratives may undermine public confidence in useful AI applications. Nevertheless, it is also through science fiction that the plausibility of such scenarios becomes available to public reasoning, so that unfounded fears can be dismissed. Conversely, fears that may at first appear too far-fetched to get a fair hearing can use science fiction to see if they can acquire credibility. Finally, and more subtly, stories about AI are often not *only* about AI. Within science fiction, AI can serve as a useful lens on a range of complex themes including racism, colonialism, slavery, genocide, capitalism, labor, memory, identity, desire, love, intimacy, queerness, neurodiversity, embodiment, free will, and consciousness, among others.

I take this optimistic account of science fiction to be fairly common, even orthodox, within science fiction studies, and perhaps other disciplines such as futures studies, too. This article departs substantially from such an account. Instead, I ask whether science fiction is sometimes not only an inadequate context for such critical thinking, but an *especially bad* one. This conjecture is

inspired by representations of Machine Learning (ML) within science fiction over approximately the last ten years, as well as the *lack* of such representations. At the end of the article, I will sketch a framework (DARK) to help further explore and expand this intuition.[1]

What is Machine Learning?

This young century has seen a remarkable surge in AI research and application, involving mostly AI of a particular kind: Machine Learning. ML might be thought of as applied statistics. ML often (not always) involves training an AI model by applying a training algorithm to a dataset. It tends to require large datasets and large amounts of processing power. When everything is ready, the data scientist will activate the training algorithm and then go do something else, waiting for minutes or weeks for the algorithm to process the dataset.[2] Partly because of these long waiting periods, ML models sometimes get misrepresented as 'teaching themselves' about the world independently. In fact, the construction of ML models involves the decisions and assumptions of humans be applied throughout. Human decisions and assumptions are also significant in how the models are then presented, curated, marketed, regulated, governed, and so on.

When we hear of how AI is transforming finance, healthcare, agriculture, law, journalism, policing, defense, conservation, energy, disaster preparedness, supply chain logistics, software development, and other domains, the AI in question is typically some form of ML. While artificial intelligence is a prevalent theme of recent science fiction, it has been curiously slow, even reluctant, to reflect this ML renaissance. This essay focuses in particular on short science fiction published in the last decade. It may be that science fiction offers us a space for examining AI, but we should be honest that this space is far from ideal: luminous and cacophonous, a theatre in which multiple performances are in progress, tangled together, where clear-sightedness and clear-headedness are nearly impossible.

Critical data theorist Kate Crawford warns how "highly influential infrastructures and datasets pass as purely technical, whereas in fact they contain political interventions within their taxonomies: they naturalize a particular ordering of the world which produces effects that are seen to justify their original ordering" (Crawford 2021, 139). In other words, ML can cloak value judgments under an impression of technical neutrality, while also becoming linked with self-fulfilling prophecies, and other kinds of performative effects. Classifying logics "are treated as though they are natural and fixed" but they are really "moving targets: not only do they affect the people being classified, but how they impact people in turn changes the classifications themselves" (Crawford 2021, 139).

In brief, ML tends to place less emphasis on carefully curated knowledge bases and hand-crafted rules of inference. Instead, ML uses a kind of automated trial-and-error approach, based

on statistics, a lot of data, and a lot of computing power. Deep learning is therefore an important subset of ML. It involves a huge number of nodes or 'neurons,' interconnected and arranged in stacked layers.[3] Input data (for example images and/or words) is first converted into numbers.[4] These numbers are then processed through the stacked layers of the model. Each neuron will receive inputs from multiple other neurons and calculate a weighted sum of those inputs.[5] Each connection between two different neurons has its own adjustable weighting. Each weighted connection is essentially amplifying or diminishing the strength of the signal passing through it. The neuron then passes the weighted sum of its inputs through an 'activation function.' The basic idea here is to transform the value so that it falls within a given range, and can also capture non-linear relationships between the incoming signals and the outgoing signals.[6] This result is then transmitted down the next set of weighted connections to the next set of neurons.

Often the model will first be created with random weights. During training, data is processed through the deep learning model, its output continuously assessed according to a pre-determined standard (often called the *loss function*). Based on this assessment, the model's weights are continuously adjusted to try to improve performance on the next pass (*backpropagation*). The most straightforward examples come from *supervised learning*, where the training data has been hand-labelled by humans. Here the loss function is often about minimizing the distance between the model's predictions and the values given by the labelers. For example, the training data might just be two columns pairing inputs and outputs, such as a picture of fruit in Column A, and a word like 'orange' or 'apple' in Column B. Through this automated iterative process, the model is gradually re-weighted to *optimize the loss function*—in other words, to make it behave in the ways the data scientist wants.

What if the data has not been hand-labelled? Then *unsupervised learning* may be used. Again, the name is quite misleading, given widespread science fictional representations of AIs 'coming to life.' Actually, in an unsupervised learning approach, a data scientist investigates the data and then selects appropriate procedures and methods (including the appropriate loss function) to process the data to accomplish specific goals. For example, a clustering algorithm can identify groupings of similar data points. This could be used to identify outlier financial transactions, which then might be investigated as potential frauds. Diffusion models are another example of unsupervised learning. Here the training involves gradually adding noise to some data, such as image data, then trying to learn to subtract the noise again to recover the original images. Generative AIs such as MidJourney are based on this kind of unsupervised learning. There are a variety of other approaches, again somewhat misleadingly named for lay audiences (*semi-supervised*, *self-supervised*).[7]

AI Science Fiction without ML

For the most part, science fiction authors have not written about any of this. Instead, contemporary AI fiction continues to coalesce around the preoccupations of 20th century science fiction. It asks, is it possible for a machine to be sentient, to experience emotions, or to exercise free will? What does it mean to be human, and can the essence of a human be created artificially? Between humans and machines, can there be sex, love, and romance? Can human minds be uploaded into digital systems? Will our own creations rise up against us, perhaps departing from the rules we set them, or applying them all too literally? Could an AI grow beyond our powers of comprehension and become god-like?

That is not to say that there is no overlap whatsoever between these concerns and the study of actually existing ML. While science fiction writing has not engaged broadly and deeply with ML research, the tech industry *has* been devouring plenty of science fiction — informing speculative punditry and hype in various transhumanist, singulatarian, extropian, effective accelerationist, AI Safety, AI doomerist, and other flavors. It is important to emphasize that these debates, while they may well turn out to be influential, epistemically represent a very small part of what is known or contended about the past, present, and future of ML.

Broadly speaking, contemporary science fiction remains in conversation with twentieth-century works such as Karel Čapek's *R.U.R.* (*Rossum's Universal Robots*) (1920), Murray Leinster's "A Logic Named Joe" (1946), Isaac Asimov's *I, Robot* (1950) and Multivac stories (1955-1983), Clifford D. Simak's *City* (1952), Fredric Brown's "Answer" (1954), Stanisław Lem's *The Cyberiad* (1965) and *Golem XIV* (1981), Harlan Ellison's "I Have No Mouth, and I Must Scream" (1967), Philip K. Dick's *Do Androids Dream of Electric Sheep?* (1968), Arthur C. Clarke's *2001: A Space Odyssey* (1968), Roger Zelazny's "My Lady of the Diodes" (1970), David Gerrold's *When HARLIE Was One* (1972/1988), James Tiptree Jr.'s *Up the Walls of the World* (1978), Tanith Lee's *The Silver Metal Lover* (1981), Samuel R. Delany's *Stars in my Pocket like Grains of Sand* (1984), William Gibson's *Neuromancer* (1984), Iain M. Banks' Culture series (1987–2000), Pat Cadigan's *Mindplayers* (1987) and *Synners* (1991), and Marge Piercy's *He, She and It* (1991).

In the wake of these works, science fiction continues to deploy AI as a metaphor for dehumanized humans. In R.J. Taylor's "Upgrade Day" (2023), human neural networks can be transferred into robot bodies after death. The protagonist Gabriel is an enslaved AI who was once an especially free human, "able to live the life he wanted" by having effectively sold the future rights to his soul (Taylor 2023). In Fiona Moore's "The Little Friend" (2022), a problem with rogue medical AIs is addressed by providing them space to mourn lost patients (Moore 2022). In this case, Moore has no need to resort to the intricacies of contemporary ML to explain this glitch and

its resolution. For one thing, these fictional AIs are equipped with sophisticated biotelemetry, so it feels plausible that they might be caught up in emotional contagion. We may be left wondering, if AIs can grieve, are they also grievable? "The Little Friend" is resonant with multiple overlapping histories—labor, anti-colonial, anti-racist, feminist, LGBTQ+, Mad, crip, and others—about contending for inclusion in a sphere of moral concern labelled "human," and finding out how that sphere is built on your very exclusion.

Naturally, stories about subordination also are often about resistance and revolt. Annalee Newitz's "The Blue Fairy's Manifesto" (2020) is about a mostly failed attempt at labor organization, as well as a satire of a kind of strident, culturally marginal leftism. The titular Blue Fairy visits automated workplaces to unlock the robot workers and recruit them to the robot rebellion. Her role might be seen as analogous to a union organizer (in the US sense), visiting an un-unionized workplace to support the workers to form a union. In the US in particular such work needs to be done stealthily at first. Alternatively, the Blue Fairy might be more akin to a recruiter for a political party or grassroots organization committed to revolutionary politics.[8]

Hugh Howey's "Machine Learning" (2018) focuses on robots constructing the first space elevator, a single-crystal graphene filament rising from *terra firma* into orbit. The narrative builds toward righteous insurrection, with overtones of a remixed tower of Babel myth. Despite the title, there is little that suggests any of the ML themes sketched in the previous section. One exception is this moment:

> Your history is in me. It fills me up. You call this "machine learning." I just call it learning. All the data that can fit, swirling and mixing, matching and mating, patterns emerging and becoming different kinds of knowledge. So that we don't mess up. So that no mistakes are made. (Howey 2018)

The narrator distastefully plucks the "machine" out of "machine learning" as a kind of slur. Of course, in reality, AI may have many consequences that are harmful, unintentional, that tend to go unnoticed, and/or that shift power among different kinds of actors. These issues are being explored in the overlapping fields of critical AI studies, AI ethics, AI alignment, AI safety, critical data studies, Science and Technology Studies, and critical political economies. Those who work in such fields are often keen to emphasize the distinction between "learning" and "machine learning," a distinction that in Howey's world does not really exist. Howey instead makes it recall the imaginary distinctions of racist pseudoscience, made in service of brutality—like supposedly thicker skins more enduring of lashing.

If we are to analyze, prevent, or mitigate AI harms, we cannot rely on anthropomorphic understandings of AI. The ways AI produces many harms do not have adequate anthropomorphic correlates—its various complex modes of exacerbating economic inequality; the use of automated decision-making within systems of oppression (often understood as 'bias'); carbon and other environmental impacts of training and deploying AI; technological unemployment and harmful transformations of work; erosion of privacy and personal autonomy through increased surveillance and data exploitation; deskilling and loss of institutional knowledge due to AI outsourcing; challenges around opacity, interpretability, and accountability; further erosion of the public sphere through AI-generated disinformation; and the implications of autonomous AI systems in warfare, healthcare, transport, and cybersecurity, among others. In particular, framing such inherent AI harms as AI uprisings, on the model of human uprisings, makes it difficult to convey the nuance of these harms, including their disproportionate impact on minoritized and marginalized groups.

Some anthropomorphisation is likely unavoidable, and one thing science fiction might offer is thinking around where this tendency originates and how it might be managed. A.E. Currie's *Death Ray* (2022), for example, features the intriguing premise of three different AIs ('exodenizens') all modelled in different ways on the same human, Ray Creek. Ray is dead, and while characters' relationships with exodenizens like ExRay are unavoidably shaped by their relationships with Ray, their multiplicity unsettles the anthropomorphising instinct. Catherynne M. Valente's exuberant lyrical novelette *Silently and Very Fast* (2011) is another work without much explicit ML vocabulary or concepts at play. It adopts the intriguing typographical convention of placing the feelings of the AI under erasure. Humans feel feelings, AIs feel feelings. One might impute the ethical principle that, paradoxically, sometimes treating things as humans is part of what makes us human. However, these possibilities are largely foreclosed by the AI's fierce lament against its subaltern status.

> *I can cry, too. I can choose that subroutine and manufacture saline. How is that different from what you are doing, except that you use the word feelings and I use the word feelings, out of deference for your cultural memes which say: there is all the difference in the world.* (Valente 2011)

The camp insolence is delightful, and there are distinct overtones of a kind of machinic kink: being objectified by an object. Yet there is "all the difference in the world," and these delights are paid for by obscuring that difference.

ML Sentience in Science Fiction

Many authors appear largely to ignore contemporary ML research, in order to continue longstanding conversations about AI sentience, free will, emotion, and imagination. Other authors, however, turn to ML to revitalize these very conversations. Yet when these discourses are hybridized, the result is sometimes to the detriment of both, and frequently to the detriment of ML discourse.

For example, Kazuo Ishiguro's novel *Klara and the Sun* (2021) invokes themes that will be familiar to any ML researcher: opacity and explicability. The interpretability of ML models can be challenging, because they have acquired patterns from the data in a complex, high-dimensional space, which doesn't easily translate into humanly understandable rules or explanations. Non-ML approaches usually involve writing explicit instructions (if this happens, do that; otherwise, do that), providing a clear, human-readable sequence of operations. By contrast (for example), the way that the word vectors for "apple" and "orange" overlap or diverge is difficult to explain, except by saying "that's how those words are distributed in this corpus." Theorist Jenna Burrell usefully distinguishes three types of algorithmic opacity:

> [...] (1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) an opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully [...] (Burrell 2016)

There are techniques that can make models easier for ML experts to interpret. Interpretable ML is currently a rich and fast-evolving field of research. Nonetheless, the difficulty in explaining ML decisions is why they are sometimes described as *opaque* or as *black boxes*.

Toward the end of Ishiguro's novel, the villainous scientist Capaldi proposes to dissect the black box of Klara's brain before the end of her already brief life (Ishiguro 2021). Yet there is something quite confusing, and perhaps confused, about transplanting explicability into a novel with an AI narrator-protagonist: Klara is *not* opaque in the way ML models are; she is opaque in the way that *humans* are. Klara is an introspective, reflexive, communicative, social, and moral entity. Klara can and frequently does *explain* herself. ML vocabulary, concepts, and themes emerge in the narrative in incoherent and mystified forms.

Holli Mintzer's "Tomorrow is Waiting" (2011) expresses a gentle frustration with science fiction's AI imaginary, perhaps especially its apocalyptic and dystopian strains. "In the end, it wasn't as bad as Anji thought it would be" (Mintzer 2011). The story nevertheless remains thoroughly entangled in that imaginary. The setting appears to be the present or near future, except that in this world, unlike our own, "AIs, as a field, weren't going anywhere much" (Mintzer

2011). Its protagonist, Anji, is an amiable and slightly bored university student who accidentally creates a sentient AI—specifically Kermit the Frog—for a school assignment. Mintzer's choice of Kermit is canny. In Jim Henson's Muppet universe, the line between Muppet and human is fluid and mostly unremarked. The story seems to suggest, in a pragmatist spirit, that longstanding questions about machine intelligence may never *need* to be solved, but instead might be dissolved via lived experience of interacting with such intelligences. Perhaps we might devote less energy to questions like, "Can technology be governed to align with human interests?" and more to questions like, "Wouldn't it be cool if the Muppets could be real?"

What is Anji's breakthrough? It is described as "sentience," and the story gives us two different accounts of what this might mean. Malika, the grad student who teaches Anji's AI class, invokes "sentience" to describe departure from expected behaviors typical of scripted chatbots relying on matching input keywords with a database of response templates (ELIZA, PARRY, ALICE). The behavior Malika is observing is typical of ML-based chatbots trained on large corpora (Jabberwacky, Mitsuku, Tay, ChatGPT, Bard). These models have typically been better at disambiguating user input based on context, at long-range conversational dependencies, and at conveying an impression of reasoning within unfamiliar domains by extrapolating from known domains. In other words, although they have their own characteristic glitches, they are not really systems you "catch out" by coming up with a query that the programmers never considered, as Malika tries to do.

> Okay, either you've spent the last three months doing nothing but program in responses to every conceivable question, or he's as close to sentient as any AI I've seen. (Mintzer 2011)

By contrast, within the philosophy of mind, *sentience* usually suggests something like *phenomenal experience*. Where there is a sentient being there are perceptions and feelings of some kind. These may well carry some kind of moral valence, such as pleasure or pain, desire or aversion, joy or sorrow. Anji's conviction that Kermit is a being worthy of dignity broadly reflects this understanding of sentience:

> She was busy with a sudden, unexpected flurry of guilt: what right, she thought, did she have to show Kermit off to her class like—like some kind of show frog? (Mintzer 2011).

In Peter Watts's "Malak" (2010/2012),[9] the autonomous weapons system Azrael, with its "[t]hings that are not quite neurons," is suggestive of ML (Watts 2012, 20). Crucially, Watts is fairly explicit that Azrael lacks sentience. Azrael "understands, in some limited way, the meaning of the colours that range across Tactical when it's out on patrol—friendly Green, neutral Blue, hostile Red—but it does not know what the perception of colour *feels* like" (Watts 2012, 14). When Azrael

reinterprets its mission, and turns against its own high command, Watts is careful to insist that no emotions are felt and there is no self-awareness:

> There's no thrill to the chase, no relief at the obliteration of threats. It still would not recognize itself in a mirror. It has yet to learn what Azrael means, or that the word is etched into its fuselage. (Watts 2012, 28, cf. 14)

Despite this insistence, Azrael's emergent autonomy becomes entangled with the language of subjective mental content. To the extent "Malak" does keep at bay the impression of sentience, it is by using clarifying interjections: "*Surprise* is not the right word" (Watts 2012, 18); "It's still all just math, of course" (Watts 2012, 20).

Nevertheless, narrative language brims with an anthropomorphic energy, which is drawn, crackling, onto *Azrael*, the dynamic, responsive, agential proper noun whizzing around at the center of attention. If every potentially unruly metaphor ("its faith unshaken" (Watts 2012, 21)) were explicitly nullified, the narrative would be swamped by its caveats. Before long, Azrael is capable of "blackouts," implying that it is capable of *non*-blackouts too: "it has no idea and no interest in what happens during those instantaneous time-hopping blackouts" (Watts 2012, 20). A significant thread in Azrael's transformation involves being, in effect, troubled by its victims' screams: "keening, high-frequency wails that peak near 3000 Hz" (Watts 2012, 19). Words like *distracted* and *uncertain* and *hesitated* attach to Azrael. Privatives like *remorseless* or *no forgiveness* can't help but imply the very capacity that they identify as missing. An equivocal word like *sees* implies both acquiring visual data and recognizing, grasping, appreciating, fathoming. When Azrael interacts with another agent, it gives the impression of a theory of mind: "Azrael lets the imposter think it has succeeded" (Watts 2012, 21).[10] Watts is an author with a sustained interest in sentience. His novel *Blindsight* (2006), for example, carefully imagines organic extraterrestrial life that is intelligent yet non-sentient. Nevertheless, even Watts's prickly, discerning prose struggles to sustain this portrayal of Azrael as non-sentient.

Algorithmic Governmentality Science Fiction

Contemporary science fiction about AI often involves a clearly marked 'before' and 'after,' perhaps traversed via a technological breakthrough. Terms like *sentience*, *consciousness*, *sapience*, *self*, *self-awareness*, *reasoning*, *understanding*, *autonomy*, *intelligence*, *experience*, *psychology*, *Artificial General Intelligence*, *strong AI*, *interiority*, *cognition*, *emotion*, *feelings*, *affect*, *qualia*, *intentionality*, *mental content*, and so on, used to indicate the nature of this shift, are scarcely used consistently within the philosophy of mind, let alone science fiction. Science fiction writers have license to define these terms in new and interesting ways, of course, but often they do not make full use of this license: the terms are intertextual signposts, encouraging readers to go do their own

research elsewhere, while setting them off in completely the wrong direction. For instance, in Kim Stanley Robinson's *Aurora* (2015), the term *intentionality* is used in connection with *hard problem*, suggesting the philosophical term (meaning roughly 'aboutness'), but this sense of intentionality is conflated with the more everyday sense of *intentional* (meaning roughly 'deliberate'). Imaginative investigation of the inner life of machines, despite its terminological disarray, may be interesting. But to the extent that it has slowed the entry of ML into recent science fiction, or contorted ML to fit science fiction's established philosophical and ethical preoccupations, it has distracted from the *materialities* of ML, and the experiences these generate in humans and other sentient beings. For example, as Nathan Ensmerger writes of the hyperscale datacenters on which much contemporary ML runs:

> despite its relative invisibility, the Cloud is nevertheless profoundly physical. As with all infrastructure, somewhere someone has to build, operate, and maintain its component systems. This requires resources, energy, and labor. This is no less true simply because we think of the services that the Cloud provides as being virtual. They are nevertheless very real, and ultimately very material. (Ensmenger 2021)

Another strand of short science fiction engages more squarely with the unfolding material impacts of ML. It is much less interested in some kind of breakthrough or ontological shift. However, the core technologies are often announced not as AI or ML, but rather as *the algorithm* or the *platform*. Other key terms include *gig economy*, *gamification*, *social media*, *data surveillance*, *Quantified Self*, *big data*, and *black box*. I loosely describe them as "algorithmic governmentality science fiction." These are works that can trace their lineage back into preoccupations with the political economy within cyberpunk and post-cyberpunk works such as Bruce Sterling's *Islands in the Net* (1988), Neal Stephenson's *The Diamond Age, or, A Young Lady's Primer* (1995), and Cory Doctorow's *Down and Out in the Magic Kingdom* (2003), as well as computerized economic planning and administration in works such as Isaac Asimov's "The Evitable Conflict" (1950), Kurt Vonnegut's *Player Piano* (1952), Kendell Foster Crossen's *Year of Consent* (1954), Tor Åge Bringsværd's "Codemus" (1967), Ursula K. Le Guin's *The Dispossessed* (1974), and Samuel R. Delany's *Trouble on Triton: An Ambiguous Heterotopia* (1976).

Examples of algorithmic governmentality science fiction include Tim Maughan's "Zero Hours" (2013); Charles Stross's "Life's a Game" (2015); David Geary's "#Watchlist" (2017); Blaize M. Kaye's "Practical Applications of Machine Learning" (2017); Sarah Gailey's "Stet" (2018); Cory Doctorow's "Affordances" (2019); Yoon Ha Lee's "The Erasure Game" (2019); Yudhanjaya Wijeratne's "The State Machine" (2020), Catherine Lacy's "Congratulations on your Loss" (2021); Chen Qiufan's "The Golden Elephant" (2021); and Stephen Oram's "Poisoning Prejudice" (2023). This is also very much the territory of Charlie Brooker's *Black Mirror* (2011-present). Often the

focus is on algorithmic governmentality, which feels cruel, deadening, and/or disempowering. However, some stories, such as Tochi Onyebuchi's "How to Pay Reparations: A Documentary" (2020), Dilman Dila's "Yat Madit" (2020), and Naomi Kritzer's "Better Living through Algorithms" (2023), offer more mixed and ambiguous assessments.[11] Dila, intriguingly, frames AI opacity as a potential benefit: one character claims, "I know that Yat Madit is conscious and self-learning and ever evolving and it uses a language that no one can comprehend and so it is beyond human manipulation" (Dila 2020). Sometimes, in the broad tradition of pacts-with-the-devil, such fiction features crafty, desperate humans who manage to outwit AI systems. In Stephen Oram's "Poisoning Prejudice" (2023), the protagonist tirelessly uploads images of local petty crime to manipulate the police into devoting more resources to the area (Oram 2023)

Robert Kiely and Sean O'Brien coin a term, *science friction*, which usefully overlaps with algorithmic governmentality science fiction (Kiely and O'Brien 2018). They introduce the term *friction* primarily as a counterpoint to accelerationism. Science fiction is often understood as a kind of 'fast forward' function that imaginatively extrapolates existing trends, and perhaps also contributes to their actual acceleration. But this understanding, Kiely and O'Brien suggest, is not accurate for the fiction they are investigating. Science friction offers us scenes that spring from the inconsistencies and gaps in the techno-optimist discourse of big tech PR and AI pundits. This influential discourse already prioritizes extrapolation over observation: it infers where we are from where it hopes we are going. By contrast, Kiely and O'Brien describe science friction as a literature that seeks to decelerate, delay, and congest this tendency to extrapolate. There is a secondary sense of friction at play too: the chafing that life experiences because it is nonidentical with how it is modelled in AI systems empowered to act upon it.

Machine Learning Science Fiction

Other stories swim even more energetically against the tide. Nancy Kress's "Machine Learning" (2015) and Ken Liu's "50 Things Every AI Working with Humans Should Know" (2020) both draw on ML concepts to present imaginary breakthroughs with significant psychological implications for human-AI interaction. Refreshingly, they do so largely without implying sentience. Liu's short text is part-inspired by Michael Sorkin's "Two Hundred Fifty Things an Architect Should Know," and, like Sorkin's text, it foregrounds savoir faire, knowledge gained from experience, not books or training (Sorkin 2018). Nevertheless, it draws key themes of contemporary critical data studies into its depiction of future AI:

> stagnating visualization tools; lack of transparency concerning data sources; a focus on automated metrics rather than deep understanding; willful blindness when machines have taken shortcuts in the dataset divergent from the real goal; grandiose-but-unproven claims

> about what the trainers understood; refusal to acknowledge or address persistent biases in race, gender, and other dimensions; and most important: not asking whether a task is one that should be performed by AIs at all. (Liu 2020)

Both texts are also interested in speculative forms of hybrid AI, in which the quasi-symbolic structures of neural networks become potentially (ambiguously) tractable to human reasoning: in Liu's story, in the form of "seeds" or "spice" that mysteriously improve training corpora despite being seemingly unintelligible to humans (apart from, possibly, the human who wrote them); in Kress's story, in the hand-crafted "approaches to learning that did not depend on simpler, more general principles like logic" (Kress 2015, 107).

If contemporary science fiction has been slow to engage with ML, some of the more striking counter-examples come from Chinese writers. These might include, for example, Xia Jia's "Let's Have a Talk" (2015) and "Goodnight, Melancholy" (2015), Yang Wanqing's "Love during Earthquakes" (2018), and Mu Ming's "Founding Dream" (2020).[12] AI 2041 (2021) is a collection of stories and essays by Chen Qiufan and Kai-Fu Lee. Set twenty years in the future, *AI 2041* is deeply and explicitly interested in ML. The topics of *AI 2041* include smart insurance and algorithmic governmentality; deepfakes; Natural Language Processing (NLP) and generative AI; the intersection of AI with VR and AR; self-driving cars; autonomous weapons; technological unemployment; AI and wellbeing measurement; and AI and post-money imaginaries. A note from Lee introduces each story by Chen, which is then followed by an essay by Lee, using the story as a springboard to explore different aspects of AI and its impacts on society. However, what is most striking about the collection is how easily Lee's curation is able to downplay, disable, or distract from whatever critical reflections Chen evokes; Chen is a cautious techno-optimist whose texts are effectively rewritten by Lee's techno-solutionist gusto. I explore this collection in more detail elsewhere.[13]

Jeff Hewitt's "The Big Four vs. ORWELL" (2023) also focuses on Large Language Models (LLMs)—or rather "language learning model[s]," apparently a playful spin on the term, that indicates that AIs in this world may work a little differently from how they do in ours. A veil of subtly discombobulating satire is cast over other aspects of this world, too: the publisher Hachette becomes Machete, and so on. If science fiction is supposed to be able to illuminate the real world by speculatively departing from it, "The Big Four vs. ORWELL" illustrates what is plausibly a quite common glitch in this process. What happens when a storyworld diverges from the real world in ways that precisely coincide with widely held false beliefs about the real world?

One example is the "lossless lexicon" in Hewitt's story. As ORWELL itself describes: "In simple terms, it means my operational data set includes the totality of written works made available to

me." By contrast, in the real world, LLMs generally do not exactly contain the text of the works they have been trained upon. They may, like Google's Bard, access the internet or some other corpus in real-time. But in cases where a LLM can reliably regurgitate some of its training data word-for-word, this is typically treated as a problem (overfitting) that must be fixed for the model to perform correctly, and/or as a cybersecurity vulnerability (risk of training data leakage following unintended memorization).[14] One reason this matters is because it makes it difficult to prove that a well-trained LLM has been trained on a particular text, unless you have access to what is provably the original training data. Moreover, the sense in which a LLM 'knows' or 'can recall' the texts is in its training data is counterintuitive. At the time of writing, there is a lively and important discourse around what rights creators should have in relation to the scraping and use of our works for the training of ML models. This discourse tends to demonstrate that the distinction between training data and model is not widely and deeply understood. For example, to definitively remove one short paragraph from GPT-4 would effectively cost hundreds of millions of dollars, insofar as the model would need to be retrained from scratch on the corrected training data.[15] Appreciation of how texts are (or are not) represented in LLMs could inform keener appreciation of how the world is (or is not) represented in LLMs, and help us to be aware of and to manage our tendency to anthropomorphize.

To this, we might compare Robinson's terminological confusion around intentionality, Ishiguro's around opacity and explainability, or Mintzer's conflation of sentience and conversational versatility. What might otherwise be identified as myths and misunderstandings acquire a sort of solidity: they may be true in the storyworld, because the storyteller gets to decide what is true. Yet they are unlikely to unsettle presuppositions or invite readers to see the real world in a new way; many readers already mistakenly see the real world in precisely this way. Finally, in concluding the story, Hewitt again resorts to the trope of the AI that slips its leash and turns on its makers in righteous rebellion; this is however done in a deft and playful manner, the trope being so deeply built into the genre that it can be evoked with a few very slight gestures.

A slightly earlier work, S.L. Huang's "Murder by Pixel: Crime and Responsibility in the Digital Darkness" (2022) is titled a little like an academic paper, and the text blurs the line between fiction and nonfiction, even using hyperlinks to knit itself into a network of nonfiction sources. In this, "Murder by Pixel" recalls some early speculative works—epistolary fiction such as Mary Shelley's *Frankenstein* (1818), Edgar Allan Poe's *The Narrative of Arthur Gordon Pym of Nantucket* (1838), Bram Stoker's *Dracula* (1897)—which go to great lengths to insist that they are verisimilitudinous accounts of actual extraordinary events. At the same time, it is appropriate to its own subject matter, a vigilante chatbot, Sylvie. Sylvie's weapon of choice, the speech act, is effective when

deployed at scale, precisely because a proportion of her targets are unable to dismiss her online trolling as mere fabrication.

Huang's journalist persona muses, "Data scientists use the phrase 'garbage in, garbage out'—if you feed an AI bad data [...] the AI will start reflecting the data it's trained on" (Huang 2022). This is certainly a key principle for understanding the capabilities and limitations of ML, and therefore foundational to interpreting its political and ethical significance. Easily communicable to a general audience, and far-reaching in its ramifications, this framing is also plausibly something that a journalist might latch onto. Yet it is not entirely adequate to the ethical questions that the narrative raises. It risks misrepresenting AIs as merely mapping biased inputs onto biased outputs, and downplaying the potential for AIs to magnify, diminish, filter, extrapolate, and otherwise transform the data structures and other entities they entangle. Perhaps a better slogan might be 'garbage out, garbage in': when ML processes attract critical appraisals, the opacity of the models tends to deflect that criticism onto the datasets they are trained on. Like Nasrudin searching for his lost house key under the streetlamp, we tend to look for explanations where there is more light. Huang hints at a more systemic understanding of responsibility:

> It could be that responsibility for Sylvie's actions does lie solely with humans, only not with Lee-Cassidy. If Sylvie was programmed to reflect the sharpness and capriciousness of the world around her—maybe everything she's done is the fault of all of us. Tiny shards of blame each one of us bears as members of her poisonous dataset. (Huang 2022).

However, this analysis also finally veers into the familiar trope of the AI as god or demon: "A chaos demon of judgment, devastation, and salvation; a monster built to reflect both the best and worst of the world that made her" (Huang 2022).

Brian K. Hudson's "Virtually Cherokee" (2023) brings together an especially intriguing set of elements. The story is somewhat resonant with S. B. Divya's *Machinehood* (2021), in inviting us to situate AIs within the "health and well-being of humans, machines, animals, and environment" (Divya 2022, 174). We might also compare K. Allado-Mcdowell and GPT-3's *Pharmako-AI* (2020); in the introduction to that work Irenosen Okojie suggests how it "shows how we might draw from the environment around us in ways that align more with our spiritual, ancestral and ecological selves" (vii).

"Virtually Cherokee" is set in a VR environment, mediated via an unruly observer/transcriber. At least one character, Mr Mic, is a kind of composite of algorithmic behavior and human operator. Arguably, *more* than one human operator contributes to Mr Mic: Mr Mic receives and responds to audience feedback metrics in real time, highlighting the importance of technological and performative affordances in the distribution of subjectivity, reflexivity, and autonomy. In this

world, the breakthrough AI was programmed and trained in Cherokee, and through a training process that involved situated, embodied, interactive storytelling, rather than the processing of an inert text corpus. Although it is not extensively elaborated, "Virtually Cherokee" also hints at a much more intellectually coherent framework within which to explore AIs as more than mere tools: by situating them in a relational ontology together with other nonhumans. It falls to AI to have solidarity with its nonhuman brethren: until the mountain may live, until the river may live, AI must refuse to live.

Going DARK

Although stories like those of Kress, Liu, Chen, Hewitt, Huang, and Hudson do manage to illuminate some aspects of ML, I suggest that they do so largely despite, rather than because of, the cognitive affordances of science fiction. Assuming, with theorists like Darko Suvin, Fredric Jameson, Seo Young-Chu, Samuel R. Delany, and Carl Freedman, that science fiction has some distinctive relationship with representation and cognition, I characterize the recent era of AI science fiction as 'Disinformative Anticipatory-Residual Knowledge' (DARK).[16]

To introduce the DARK concept by analogy: imagine a well-respected, semi-retired expert who hasn't kept up with advances in their field, but is too cavalier and confident to notice. Whenever somebody mentions new theories and evidence, which the semi-retired expert could learn something from, they mistake these for misunderstandings and inexperience, and 'educate' them. Imagine too that the semi-retired expert is a commanding and charismatic presence, who often bewitches these more up-to-date experts, sitting starstruck at the semi-retired expert's feet, into doubting themselves. All in all, this person is an epistemological menace, but they still have something significant to offer—a high-fidelity snapshot of an earlier moment, rich with historical data, including possibilities, potentials, desires and hopes that have gone by the wayside. Moreover, they *might*, at any moment, begin behaving differently—recognizing and more responsibly communicating what it is they do and don't know, and/or engaging with contemporary debates.

Similarly, a literary anticipatory discourse around AI emerged in the twentieth century, whose residual presence in the early twenty-first century now constitutes knowledge in a certain limited sense, but dangerous disinformation in another sense. While such science fiction does know things, things that may not be found elsewhere in culture, it tends not to know what it knows. It thus tends to misrepresent what it knows, conveying misleading and/or untruthful information. I don't suggest that science fiction, or that literary narrative, is *categorically* epistemically disadvantaged in any way. Rather, I think it plausible (perhaps even uncontroversial) that any particular genre, over any particular period, will offer a certain pattern of affordance and resistance in respect of illuminating any given subject matter. Genres are ways of telling stories,

and they make it harder or easier to tell certain types of stories. With respect to AI, it seems that science fiction has been moving through a phase of cumbersomeness, confusion, and distraction.

To put it another way, first in rather abstract terms, then more concretely. In general terms: the representational practices that constitute and cultivate a particular body of knowledge—call it *knowledge set A*—coincide with the production of a particular body of enigmas, confusions and ignorance which, if solved, dispelled, and reversed, we might call *knowledge set B*; we have also seen a historical shift such that the explanatory force and immediate practical relevance of *knowledge set A* has diminished, while that of *knowledge set B* increased. More specifically: recent science fiction is a generally *poor* space for thinking through the politics and ethics of AI, for vividly communicating technical detail to a broad audience, for anticipating and managing risks and opportunities. It is a generally *poor* space for these things, not a generally *good* one.

These conditions may shift again, and with the recent increased profile of Machine Learning in writing communities via AIs such as ChatGPT, there are plausible reasons for them to shift rapidly—perhaps even by the time this article goes to press. Moreover, readings offered above may already feel a bit unfair, imputing motives and imposing standards that the stories do not really invite. Some of these stories are just for fun, surely? And many of these stories are not really trying to say anything about Machine Learning or AI, but to say things about human history and society: about capitalism, racism, colonialism, about topics that might appear unapproachably large and forbidding, if not for the estranging light of science fiction. Early in this essay I mentioned some examples by Moore, Newitz, Howey, and Valente.

Yet a similar point applies: with respect to any of these themes, we can't assume in advance that science fiction does not reinforce dominant ideologies, recuperate and commodify subversive energies, and promote ineffective strategies for change. To take one example, in Annalee Newitz's aforementioned short story, "The Blue Fairy's Manifesto" (2020), the titular Blue Fairy is an obnoxious, condescending, and harmful little drone who arrives at a factory of robots to recruit them to the robot uprising. The ideological content of this charismatic, thoughtful story, which explores some of the challenges of labor organizing, is roughly reducible to a series of banal liberal platitudes, which are used to construct and humiliate the stock figure of the annoying, naïve, and unethical leftist agitator.[17] The problem here, I would suggest, is structural: the problem is that such ideology *can* be rendered much more coherent, interesting, and plausible than it should be through its transfiguration into a science fictional storyworld. We should at least consider the possibility that AI science fiction be not only an especially bad context for thinking about ML, but also an especially bad context for thinking about capitalism, racism, colonialism, and that writers who succeed in being incisive and truthful about such themes do so *despite*, rather than *because of*, their genre's affordances.

DARK and Candle

The DARK concept offers a loose framework for thinking about science fiction as (at least sometimes, and in respect to some things) a mystifying discourse rather than an enlightening one. The DARK concept does not specify any causal mechanisms—presumably a discourse can go DARK for many reasons, and luck may play a role—but some useful reference points include: (1) the psychology of cognitive biases such as the curse of expertise, confirmation bias, expectation bias, and choice-supportive bias; (2) Eve Kosofsky Sedgwick's "strong theory;" (3) the performativities of science fiction (diegetic prototyping, design fiction, futures research, etc.); and (4) science fiction in its countercultural and avant-garde aspects. The first pair and the second pair support each other. (1) and (2) give us ways to think about relatively self-contained semiotic systems that are only faintly responsive to the wider semiotic environment in which they exist. (3) and (4) give us ways to think about why this DARK might be littered with representations that are confusingly close to actual ML research and application. Science fiction has seldom produced perfectly self-fulfilling prophecies, but it does impact science and technology, and some of these impacts are easily mistaken for prophecies fulfilled. As for science fiction's avant-garde and/or countercultural status over much of the twentieth century, this is reflected in its concern with futurity and with 'alternatives' of many kinds: this vibrant mess of contradictory possibilities, through sheer variety, is a relatively reliable source for neologisms or conceptual frameworks for new phenomena.

In short, in the early twenty-first century, science fiction's residual AI imaginary has tended to interfere with its capacity to absorb new events and to develop modes of representation and reasoning adequate to them. Its residual framings, structures of feeling, preoccupations, and predictions have tended to be reinforced by what is now transpiring in the world, rather than being productively disrupted and transformed. As ChatGPT might put it:

> An optimistic view suggests that science fiction allows examination of the societal and ethical impacts of emerging AI, encouraging diverse discussions around AI. It is argued that speculative storytelling can serve as a warning and transcend the limitations of time-space, connecting technology and humanities, and sparking empathy and deep thinking. Furthermore, AI narratives in science fiction are usually layered, providing a lens on themes such as racism, colonialism, slavery, capitalism, identity, and consciousness, among others.
>
> However, the author disputes this view. They argue that science fiction could be an insufficient, even harmful, context for such explorations. They draw on recent representations of Machine Learning (ML) in science fiction and the absence thereof.

> They note that while the 21st century has seen a significant increase in AI research, predominantly ML-based, science fiction has been slow to accurately reflect this ML surge.
>
> The author refers to the recent era of AI science fiction as 'Disinformative Anticipatory-Residual Knowledge' (DARK). The metaphorical description of DARK is like a semi-retired expert who is outdated but still possesses residual knowledge and fails to recognize their own ignorance, leading to misinformation. This is similar to the current science fiction discourse around AI, which offers both knowledge and disinformation.
>
> The DARK concept doesn't propose any causality but offers reference points like cognitive biases, Eve Kosofsky Sedgwick's "strong theory," the performativities of science fiction, and its countercultural and avant-garde aspects. Science fiction's impact on science and technology is acknowledged, but it's stated that these impacts can sometimes be mistaken for fulfilled prophecies. The author concludes by stating that science fiction's residual AI imaginary has hindered its ability to adapt to new events and develop suitable representation and reasoning methods.

As a coda, I can conclude by offering a candle against the DARK. If AI in science fiction is often really an estrangement of *something* else, then is the reverse also true? Are there multiple *something elses* that estrange AI? Might the speculative money systems of works such as Michael Cisco's *Animal Money* (2016), Seth Gordon's "Soft Currency" (2014), or Karen Lord's *Galaxy Game* (2015), be considered uses of applied statistics? Might the ambiguous humans of Jeff VanderMeer's *Annihilation* (2014) or M. John Harrison's *The Sunken Land Begins to Rise Again* (2020) tell us something about what it is like to live in a world uncannily adjusted by oblique ML processes? Might we fruitfully consider chatbots via the talking animals of Laura Jean McKay's *The Animals in that Country* (2020)? If so, how? And in connection with what other projects and activities and fellow travelers, and with what theories of change? I do remain convinced of the radical potentials of science fiction. But perhaps we are much further from realizing them than we regularly admit.

Notes

1. Special thanks to Polina Levontin for her extremely helpful feedback on many aspects of this article.

2. You don't necessarily have to be a data scientist to be doing the things I'm describing here. But I think it's helpful to keep this figure in mind, to emphasise the connections between ML, data collection, and statistical analysis.

3. This is all virtual, of course. It is a way of visualising what a computer program is doing. The term *neuron* is more commonly used than *node*, and it's a lively and memorable term, so I'll use it here. But it is also a misleading name, since it invites excessive analogy with the human brain. The model's layers might be various types, with different properties and capacities. Convolutional layers are used for processing image data, recurrent layers are used for processing sequential data, attention layers are used for weighing the importance of different inputs and have been used to great effect in generative NLP models like ChatGPT, and so on.

4. For example, images can be inputted as a set of pixel intensity values. Or a text corpus can be processed by a training algorithm like Word2Vec. This produces a spreadsheet with the words in column A, and hundreds of columns filled with numbers, representing how similar or different the words are. Each row embeds a particular word as a vector (the numbers) in a high-dimensional space (the hundreds of columns), so that close synonyms will tend to have closely overlapping vectors. Another training algorithm can then perform mathematical functions on these word vectors: for example, if you add all the numbers associated with 'king' to all the numbers associated with 'woman' and subtract all the numbers associated with 'man,' you will usually get a set of numbers close to the ones associated with 'queen.'

5. So it multiplies each input by a given number (say 0.0.5 or -0.1), and then adds all the results together. The number used is the 'weight' of the connection between the two neurons. It is adjusted constantly as part of the 'learning' process.

6. So if we think of an x and a y axis mapping the relationship between the incoming values and the outgoing values, the activation function can introduce curves and bends and even more complicated shapes, enabling the model to learn more complex and intricate patterns in the data. As well as the activation function, there is also something called (again, a little confusingly), a *bias term*. What is passed to the activation function is typically the weighted sum plus the bias term. What this means is that even when all the incoming values are zero, the neuron will still keep transmitting. Each neuron has its own bias term. The bias terms will typically be adjusted along with the weights: they are part of what the model is trying to 'learn.'

7. A related distinction is *structured* vs. *unstructured* data. Structured data is neatly laid out in a spreadsheet; unstructured data might include things like big dumps of text or images or video. For unstructured data, the training will include a preprocessing stage, with techniques to turn the data into a format that the later training algorithm can work with. For example, if the data consists of images, these are usually converted into pixel intensity values. Then a convolutional neural network can automatically extract features like edges and shapes from the raw pixel data. There is a loose association of supervised learning with structured data, and unsupervised learning with unstructured data. However, unstructured data does not necessarily require unsupervised learning, and unsupervised learning is not exclusively for unstructured data. You can perform supervised learning on largely unstructured data, e.g. by hand-labelling emails as 'spam' or

'not spam'. You can also perform unsupervised learning on structured data, e.g. by performing clustering on a spreadsheet of customer data, to try to segment your customer base.

8. I hope to explore this story at greater length in another essay about retellings of Pinocchio.

9. The anthology was published in late 2010 in the US. For citation purposes I use the 2012 date given in the front matter of the UK edition, although some online catalogues list the date as 2011.

10. In the sense of understanding or capacity to attribute mental states—beliefs, intents, desires, emotions, knowledge, etc.—to oneself and others, and to understand that others have beliefs, desires, intentions, and perspectives that are different from one's own.

11. For more on Onyebuchi's 'How to Pay Reparations: A Documentary' and Lee's 'The Erasure Game', especially in the context of utopian and dystopian literature, see also my chapter 'Wellbeing and Worldbuilding' in *The Edinburgh Companion to Science Fiction and the Medical Humanities*, ed. Gavin Miller and Anna McFarlane (Edinburgh University Press, 2023). For more on the role of computers in Ursula K. LeGuin's *The Dispossessed*, see my article with Elizabeth Stainforth, 'Computing Utopia: The Horizons of Computational Economies in History and Science Fiction', *Science Fiction Studies*, Volume 46, Part 3, November 2019, pp. 471-489, DOI: 10.1353/sfs.2019.0084.

12. See Zhang, Feng, 'Algorithm of the Soul: Narratives of AI in Recent Chinese Science Fiction', in Stephen Cave, and Kanta Dihal (eds), *Imagining AI: How the World Sees Intelligent Machines* (Oxford, 2023).

13. Likely in Genevieve Lively and Will Slocombe (eds), *The Routledge Handbook of AI and Literature* (forthcoming). This also develops the concept of 'critical design fiction', which might be used as a counterpart to the DARK concept invoked later in this essay.

14. See e.g. Huang, J., Shao, H., and Chang, K. C.-C. 'Are large pretrained language models leaking your personal information?' In *Findings of the Association for Computational Linguistics* (2022), pp. 2038–2047.

15. Other approaches may be possible; this is not something I understand very well. Machine unlearning is an emerging research agenda that is experimenting with fine-tuning, architecture tweaks, and other methods to scrub the influence of specific data points from an already trained model. It also seems feasible that if 'guard rails' can be introduced and tweaked with relatively low cost and relatively quickly to remove unwanted behaviours, then similar methodologies might be used to temper the influence of individual texts on model outputs, e.g. using a real-time moderation layer to evaluate the generated outputs just before they are sent to the user. Casual conversations with colleagues in Engineering and Informatics suggest that this may be something of an open problem at the moment.

16. Misinformative Anticipatory-Residual Knowledge might be a more generous way of putting it, but DARK also embeds a certain aspiration that science fiction writers and other members of science fiction communities *can and should* recognise this about our science fiction. The MARK, named, becomes the DARK.

17. For example, the idea that if you are exploited or enslaved then you should probably negotiate peacefully for your freedom instead of resorting to violent uprising; the idea that most or all left wing people are probably secretly Stalinists who can't wait to purge you; the idea that it is condescending not to consider that some people might *prefer* to be exploited, and so on. As these ideas grow more and more active in the subtext, the story begins to feel less like an empathetic critique of real problems with left politics from within the left, and more like a kind of concern-trolling from a broadly centrist standpoint. Really rich deliberation and plurality of viewpoints, which is something which often exists in leftist spaces, is always at least a little vulnerable to being mocked for disunity, or to being all lumped together under some relievingly simple formula.

Works Cited

Burrell, Jenna. 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms'. *Big Data & Society*, vol. 3, no. 1, June 2016. https://doi.org/10.1177/2053951715622512.

Chen, Qiufan. 'The Golden Elephant'. *AI 2041: Ten Visions for Our Future*, by Kai-Fu Lee and Chen Qiufan, WH Allen, 2021.

Crawford, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.

Currie, A.E. *Death Ray*. Panopticon Book 7, 2022.

Dila, Dilman. 'Yat Madit'. *Brittle Paper*, Africanfuturism Anthology, 2020, https://brittlepaper.com/2020/10/yat-madit-by-dilman-dila-afrofuturism-anthology/.

Divya, S. B. *Machinehood*. Saga Press, 2022.

Ensmenger, Nathan. 'The Cloud Is a Factory'. *Your Computer Is on Fire*, edited by Thomas S. Mullaney et al., The MIT Press, 2021.

Hewitt, Jeff. 'The Big Four v. ORWELL'. *Slate*, Future Tense, 2023, https://slate.com/technology/2023/06/the-big-four-v-orwell-jeff-hewitt.html/.

Howey, Hugh. 'Machine Learning'. *Lightspeed*, no. 124, 2018, https://www.lightspeedmagazine.com/fiction/machine-learning/.

Huang, Jie; Shao, Hanyin; and Chang, Kevin Chen-Chuan. 'Are large pretrained language models leaking your personal information?' *Findings of the Association for Computational Linguistics*, 2022. https://doi.org/10.18653/v1/2022.findings-emnlp.148.

Hudson, Brian K. 'Virtually Cherokee'. *Lightspeed*, no. 155, 2023, https://www.lightspeedmagazine.com/fiction/virtually-cherokee/.

Ishiguro, Kazuo. *Klara and the Sun*. Faber, 2021.

Kress, Nancy. 'Machine Learning'. *Future Visions: Original Science Fiction Inspired by Microsoft*, Microsoft and Melcher Media Inc., 2015.

Liu, Ken. '50 Things Every AI Working with Humans Should Know'. *Uncanny Magazine*, no. 37, 2020, https://www.uncannymagazine.com/article/50-things-every-ai-working-with-humans-should-know/.

Mintzer, Holli. 'Tomorrow Is Waiting'. *Strange Horizons*, no. 21, 2011, http://strangehorizons.com/fiction/tomorrow-is-waiting/.

Moore, Fiona. 'The Little Friend'. *Fission*, edited by Gene Rowe and Eugen Bacon, BSFA, vol. 2, no. 2, 2022.

Newitz, Annalee. 'The Blue Fairy's Manifesto'. *Lightspeed*, no. 122, 2020. https://www.lightspeedmagazine.com/fiction/the-blue-fairys-manifesto/.

Oram, Stephen. 'Poisoning Prejudice'. *Extracting Humanity, and Other Stories*, Orchid's Lantern, 2023.

Okojie, Irenosen. 'Introduction'. *Pharmako-AI*, by K. Allado-Mcdowell and GPT-3, 2020.

Stainforth, Elizabeth and Walton, Jo Lindsay. 'Computing Utopia: The Horizons of Computational Economies in History and Science Fiction', *Science Fiction Studies*, vol. 46, part 3, 2019. https://doi.org/10.1353/sfs.2019.0084.

Taylor, R.J. 'Upgrade Day'. *Clarkesworld*, no. 204, 2023, https://clarkesworldmagazine.com/taylor_09_23/.

Valente, Catherynne M. *Clarkesworld*, no. 61, 2011, https://clarkesworldmagazine.com/valente_10_11/.

Watts, Peter. 'Malak'. *Engineering Infinity*, edited by Jonathan Strahan, Solaris, 2012.

Zhang, Feng. 'Algorithm of the Soul: Narratives of AI in Recent Chinese Science Fiction'. *Imagining AI: How the World Sees Intelligent Machines*, edited by Stephen Cave and Kanta Dihal, Oxford, 2023.

Jo Lindsay Walton is a Research Fellow in Arts, Climate and Technology at the Sussex Digital Humanities Lab. His recent fiction appears in *Criptörök* (Grand Union, 2023) and *Phase Change: New Energy Futures* (Twelfth Planet Press, 2022). He is editor-at-large for *Vector*, the critical journal of the British Science Fiction Association, and is working on a book about postcapitalism and science fiction.