Exploring artificial intelligence futures

Shahar Avin*

Abstract

Artificial intelligence technologies are receiving high levels of attention and 'hype', leading to a range of speculation about futures in which such technologies, and their successors, are commonly deployed. By looking at existing AI futures work, this paper surveys, and offers an initial categorisation of, several of the tools available for such futures-exploration, in particular those available to humanities scholars, and discusses some of the benefits and limitations of each. While no tools exist to reliably predict the future of artificial intelligence, several tools can help us expand our range of possible futures in order to reduce unexpected surprises, and to create common languages and models that enable constructive conversations about the kinds of futures we would like to occupy or avoid. The paper points at several tools as particularly promising and currently neglected, calling for more work in data-driven, realistic, integrative, and participatory scenario role-plays.

Key words : horizon scanning, scenario planning, interactive fiction, role-play, superintelligence

* Research Associate in Centre for the Study of Existential Risk, University of Cambridge

< contents >

1. Introduction
2. Exploring artificial intelligence futures
3. Summary and conclusion

1. Introduction

"Artificial intelligence" (AI) is one of the more hyped-up terms in our current world, across academia, industry, policy and society (Shoham, Perrault, Brynjolfsson, & Clark, 2017). The interest in AI, which long predates the current fascination, has given rise to numerous tools and methods to explore the potential futures of the technology, and its impacts on human lives in a great variety of domains. While such visions are often drawn to utopian or dystopian extremes, more nuanced perspectives are also plentiful and varied, drawing on the history of the field, measurable progress and domain-specific expertise to extrapolate into possible future trends.

This paper presents a survey of the different methods available for the exploration of AI futures, from narrative fictions in novels and movies, through disciplinary expert study of e.g. economic or philosophical aspects of AI futures, to integrative, interdisciplinary and participatory methods of exploring AI futures.

I begin in this section with setting common terms and boundaries for the discussion: the boundaries of 'artificial intelligence' for the purposes of this paper, certain contemporary technologies and trends that help ground and define the space of exploration, and an outline of the utopian and dystopian extremes that bound the current imagination of AI

futures. I then go through each method of futures exploration in turn, providing a few examples and discussing some of the advantages and shortcomings of each. I conclude with a summary of the different methods and suggestions of strategies that may help furnish us with better information and expectations as we progress into a future shaped by AI.

1.1 Defining artificial intelligence

Given the newfound interest in AI, it is important to remember the history of AI as a field of research originating from work during the Second World War on computation and encryption, and the visions of the field's founders of machines that can learn and think like humans (Turing, 1950).

While a precise definition of AI is elusive, I will satisfy myself with an analogy to artificial hearts and lungs: machines that can perform (some of) the functions of biological systems, in this case the human or animal brain/nervous system, while at the same time lacking other functions and often differing significantly in shape, material and other properties; this behavioural definition coheres well with the imitation game, or Turing test, that focuses on the machine "passing as" a human in the performance of a specific, delineated task within a specific, delineated domain. As the tasks become more vague, multifaceted and rich, and the domain becomes wider and less well defined, we move on the spectrum from narrow to general intelligence (Legg & Hutter, 2007).

The history of the field of AI research shows how wrong we tend to be, *a priori,* about which tasks are going to be easy, and which will be hard, for a machine to perform intelligently (Minsky, 1988; Moravec,

1986). Breakthroughs in the field are often indexed to new exemplars of classes of tasks being successfully automated, for example game playing (Campbell, Hoane Jr, & Hsu, 2002; Silver et al, 2017) or image classification (Krizhevsky, Sutskever, & Hinton, 2012).

1.2 Contemporary artificial intelligence

The current AI hype cycle is dominated by machine learning, and in particular by deep learning (LeCun, Bengio & Hinton, 2015). Relying on artificial neural networks, which emerged as broadly neurologically inspired algorithms in the second half of the 20th century (Lippmann, 1987), these methods gained newfound success with the increasing availability of fast hardware and of large labelled datasets (Amodei & Hernandez, 2018).

In recent years we have seen increasing applications of deep learning in image classification, captioning, text comprehension, machine translation, and other domains. In essence, the statistically-driven pattern recognition afforded by these technologies presented a sharp break from previous conceptions of AI as logic/rule-based, and a transition from the domain of explicit expert knowledge to domains of split-second recognition and response tasks (including, for example, driving-related tasks). However, the revolution also touched on expert domains that rely on pattern recognition, including medical image diagnosis (Esteva et al, 2017) and Go game play (Silver et al, 2017).

Alongside these broadly positive developments, we have seen more ethically questionable applications, including in speech (Lyrebird, 2018) and video synthesis (Suwajanakorn, Seitz, & Kemelmacher-Shlizerman, 2017) that mimics existing individuals, in learning to execute cyber

attacks (Fraze, 2018), and in profiling and tracking individuals and crowds based on visual, behavioural and social patterns (Zhang, Li, Wang, & Yang, 2015). Existing and near future technologies enable a range of malicious use cases which require expanded or novel policy responses (Brundage, Avin et al, 2018).

1.3 Possible artificial intelligence futures

As we look further into the future, our imagination is guided by common tropes and narratives that predate the AI revolution (Cave & Dihal, 2018).

On the utopian end, super-intelligent thinking machines that have our interests as their guide, or with which we merge, could solve problems that have previously proven too hard to us mere humans, from challenges of environmental management and sustainability, to advanced energy sources and manufacturing techniques, to new forms of non-violent communication and new worlds of entertainment, to medical and biological advances that will make diseases a thing of the past, including the most terrifying disease of all – ageing and death (Kurzweil, 2010).

On the dystopian end, robotic armies, efficient and entirely lacking in compassion, coupled with the ability to tailor propaganda to every individual in every context on a massive scale, suggest a future captured by the power-hungry, ruthless few, with no hope of freedom or revolution (Mozur, 2018; Turchin & Denkenberger 2018).

Worse still, if we ever create super-intelligent artificial systems, yet fail to align them with humanity's best interests, we may unleash a process of relentless optimisation, which will (gradually or rapidly) make our planet an uninhabitable environment for humans (Bostrom, 2014).

The danger with extreme utopian and dystopian visions of technology futures is that they chart out what biologist Drew Endy called "the half pipe of doom" (Endy, 2014), a dynamic where all attention is focused on these extreme visions. More attention is warranted for mapping out the rich and complex space in between these extremes.

2. Exploring artificial intelligence futures

We are not mere bystanders in this technological revolution. The futures we occupy will be futures of our own making, by action or inaction. To take meaningful action, we must come prepared with a range of alternatives, intervention points, a map of powerful actors and frameworks of critique. As the technical advances increasingly become widely accessible (at least on some level), it is our responsibility, as scholars, policy makers, and citizens, to engage with the technical literature and communities, to make sure our input is informed and realistic.

While it is the responsibility of the technical community to engage audiences affected by their creation (which, in the context of AI technologies, seems to be everyone), there is also a responsibility for those in the relevant positions to furnish decision makers (again, broadly construed) with rich and diverse, yet fact-based and informed, futures narratives, maps and scenarios. Below I will survey a variety of tools available to us for exploring such futures, pointing out a few examples for each and considering advantages and limitations for each tool.

As a general note, this survey aims to be illustrative and comprehensive, but does not claim to be exhaustive. The examples chosen are by no means representative or exemplary − they are strongly biased by

my regional, linguistic and disciplinary familiarity and preferences. Nonetheless, I hope the overall categorization, and analysis of merits and limitations, will generalise across languages, regions and disciplines. I look forward to similar surveys from other perspectives and standpoints.

2.1 Fictional narratives

Probably the most widely recognised source of AI futures is fictional narratives, across different media such as print (novels, short stories, and graphic novels), music, films and television. These would often fall within the science fiction genre, or one of its numerous sub-genres. A few examples, chosen somewhat carelessly from the vast trove of AI fictions, include Asimov's *Robot* series, Leckie's *Imperial Radch* trilogy, Banks' *Culture* novels, Wells' *Murderbot Diaries* series, *The Jetsons*, the *Terminator* franchise of movies and TV series, the movie *Metropolis,* and the musical concept series of the same name by Monáe.

Works vary greatly in their degree of realism, from those rich in heavily researched details, to those that deploy fantastical technology as a tool to explore some other topic of interest, such as emotions, power relations, agency or consciousness. As such, fictional AI narratives can be both a source of broadened horizons and challenging ethical questions, but also a source of harm when it comes to exploring *our* AI futures – they can anchor us to extreme, implausible or misleading narratives, and, when they gain widespread popularity, can prevent more nuanced or different narratives from gaining attention.

The challenge for fictional AI narratives to provide useful guidance is further aggravated by four sources: the need to entertain, the pressure

to embody, a lack of diversity, and a limited accountability.

2.1.1 The need to entertain

Authors and scriptwriters need to eat and pay rent, and the amount of remuneration they receive is linked to the popularity of their creations, either directly through sales or indirectly through the likelihood of contracting. Especially with high-budget production costs, e.g. in Hollywood films (De Vany, 2004), scripts are likely to be more popular if they elicit a positive response from a broad audience, i.e. when they entertain. There is no *prima facie* reason to think that what makes for good entertainment also makes for a useful guide for the future, and many factors are likely to point to these two coming apart, such as the cognitive load of complexity and other cognitive biases (Yudkowsky, 2008), or the appeal of extremes (Kareiva & Carranza, 2018; Needham & Weitzdörfer, forthcoming).

2.1.2 The pressure to embody

Especially in visual media, but also in written form, narratives are made more accessible if the AI technologies discussed are somehow concretised or embodied, e.g. in the form of robots, androids, cyborgs or other machine bodies (Kakoudaki, 2014). Such embodiment serves as a useful tool for exploring a range of pertinent issues, but also runs the risk of distracting us from other forms of intelligence that are less easy to make tangible, such as algorithms, computer networks, swarm intelligence and adaptive complex systems. The pressure to embody relates to, and is made complicated by, the proliferation of embodied instances and fictions of artificial intelligence, either as commercial products (Harris, 2017) or as artistic creations of robots and thinking machines in visual and physical forms, for example robots toys or the

illustrations that accompany news articles and publications. In general, as per my definition in the beginning, understanding of artificial intelligence should focus on action and behaviour rather than form, though there are good arguments suggesting the two are linked (Shanahan, 2010).

2.1.3 Lack of diversity

While narrative fictions may well provide us with the most rich and diverse exploration of possible AI futures, we should be mindful that not all identities and perspectives are represented in fictional narratives, and that the mere existence of a work does not readily translate into widespread adoption; narratives, like individuals, groups and world views, can be marginalised. While science fiction has been one of the outlets for heterodox and marginalised groups to make their voices heard (Rose, 2000), this is not universally welcome (Oleszczuk, 2017), and the distribution of attention is still heavily skewed towards the most popular works (De Vany, 2004).

2.1.4 Limited accountability

Creators of fictional narratives receive feedback from two main sources, their audience (through purchases and engagement with their works) and their critics. While these sources of feedback may occasionally comment or reflect on a work's ability to guide individuals and publics as they prepare for the future, this is not seen as a main aim of the works not an essential part of it (Kirby, 2011). In particular, there is little recognition of the possible harms that can follow misleading representations, though it is reasonable to argue that such harms are limited, especially in the absence of better guidance, and the fact that experts deliberately aiming to provide such guidance tend to fare quire poorly (Armtrong & Sotala, 2015).

2.2 Single-discipline futures explorations

As part of the phenomenon of AI hype, we are seeing an increase in the number of non-fiction books exploring the potential implications of artificial intelligence for the future, though of course such books have been published since before the field became established in academia, and previous 'AI summers' have led to previous periods of increased publication. The authors who publish on the topic come from a wide range of disciplines, and deploy varying methods and arguments from diverse sources. These contribute to a richer understanding of what is, at heart, a multifaceted phenomenon.

For example, AI researchers (Boden, 2016; Domingos, 2015; Shanahan, 2015) spend just as much time on the history and sociology of the field, and on dispelling misconceptions, as they do on laying down observations and arguments with relevance for the future; mathematicians and physicists (Fry, 2018; Tegmark, 2017) focus on the world as seen through the lens of information, models and mathematics, and the AI futures that such a perspective underwrites; technologists focus on underlying technology trends and quantitative predictions (Kurzweil, 2010); risk analysts explore the various pathways by which AI technologies could lead to future catastrophes (Barrett & Baum, 2017; Turchin & Denkenberger, 2018); economists focus on the impacts of AI technologies on the economy, productivity and jobs (Brynjolfsson & McAfee, 2014; Hanson, 2016); self-published, self-proclaimed business thought-leaders share their advice for the future (Hyacinth, 2017; Rouhianien, 2018); political commentators write manifestos arguing for a particular future (Srnicek & Williams, 2015; Bastani, 2018); and philosophers examine the very nature of intelligence, and what happens when we extrapolate our understanding of it, and related concepts, into

future capabilities that exceed what evolution has been able to generate (Bostrom, 2014).

While the quality of research and arguments presented in such works tends to be high (as academic and public reputations are at stake), any predictions presented in such works tend to fair poorly, due to numerous factors including biases, partial perspectives, non-linear and discontinuous trends, hidden feedback mechanisms, and limited ability to calibrate predictions (Armstrong & Sotala, 2015; Rowe & Beard, 2018, Yudkowsky, 2017). Furthermore, disagreement between experts, while to be expected given the uncertainties involved, can have a paralyzing effect for audiences, a fact that can be exploited (Baum, 2018).

If fictional narratives are best seen as a rich and fertile ground for futures imagination (as long as we do not get too distracted by the flashy and popular), expert explorations provide a rich toolset of arguments, trends and perspectives with which we can approach the future with an informed, critical stance, as long as we appreciate the deep uncertainty involved and avoid taking any trend or prediction at face value.

2.3 Group-based futures exploration

The nature of the problem being addressed – what are possible AI futures and which ones we should aim for or avoid (and how) – is inherently complex, multi-faceted and interdisciplinary. It is therefore natural to explore this problem through utilising diverse groups. There are various methods to do this, each with advantages and disadvantages (Rowe & Beard, 2018).

2.3.1 Expert surveys

What do different individuals think about the future of AI? One way to find out is to ask them. While survey design is not an easy task, we have the ability to improve upon past designs, and regularly update our questions, the target community, and the knowledge on which they draw (as more experience is gained over time).

Surveys amongst experts have been used in particular to explore questions of timing and broad assessment of impact – when will certain capabilities become available and will they have a positive or negative impact (Grace, Salvatier, Dafoe, Zhang, & Evans, 2017; Müller & Bostrom, 2016). As surveys only tell us *what* people think, rather than *why* they think it, they are best treated not as a calibrated prediction of the future (as all estimates could be flawed in the same way), but rather a useful data point about what beliefs are prevalent right now, which in itself is useful for exploring what beliefs might hold currency in the future, and how these might affect the future of AI.

2.3.2 Public polling

Public polling aims to examine both public understanding of the technology, the desirability of possible applications and concerns about possible uses and misuses of the technology (The Royal Society, 2017). While it may be tempting to interpret these polls as “hard data” on public preferences, it should be remembered that many factors affect responses (Achen & Bartels, 2017). In the Royal Society study cited above, conducted by Ipsos Mori, poll findings were compared with surveys of focus groups that had in-depth interactions with experts and structured discussions around the survey questions. Such practices bring polling closer to participatory futures workshops, discussed below.

2.3.3 Interdisciplinary futures studies

Often we would want to go beyond an aggregate of single points-of-view, aiming for a more holistic understanding of some aspect of the future of AI through interactions between experts. Such interactions can be one-off or long standing, and they can be more or less structured (Rowe & Beard, 2018). An example of a broad-scoped, long-term academically led interdisciplinary study is the Stanford 100 year study of artificial intelligence (Grosz & Stone, 2018). An example of a more focused study is the workshop that led to the report on the potential for malicious use of artificial intelligence (Brundage, Avin et al, 2018). While such studies offer a depth advantage over surveys, and a diversity advantage over single-domain studies, they still face challenges of scope and inclusion: too narrow focus, on either topic or participants, can lead to a narrow or partial view, while too broad scoping and inclusion can make the process unmanageable (Collins & Evans, 2002; Owens, 2011).

2.3.4 Evidence synthesis and expert elicitation

With a growing evidence base relevant to AI futures, policy making and policy guiding bodies are beginning to conduct structured evidence synthesis studies (British Academy & The Royal Society, 2018). The methodologies for conducting such studies have been improved over the years in other evidence-reliant policy domains, and many lessons can be ported over, such as making evidence synthesis more inclusive, rigorous, transparent and accessible (Donnelly et al, 2018; Sutherland & Worldley, 2018).

We are also seeing efforts from governments to solicit expertise from a broad range of source, as early fact-finding steps that could lead to or inform policy in this space (Felten & Lyons, 2016; House of Lords,

2018). While such efforts are welcome, both in their interdisciplinary and participatory nature, through their democratic mandate, and through the proximity of expertise and accountable decision making, it should be noted that results still very much depend on the experts in the room, that such exercises tend to avoid areas of high uncertainty or disagreement (which may be the areas demanding most attention), and that the issues are often global and open in nature, limiting the effectiveness of national strategy and regulation.

2.4 Extrapolating from past and current data trends

While historical trends may provide only a limited guide to the future when it comes to emerging technologies (Farmer & Lafond, 2016), it is still useful to have an up-to-date understanding of the state-of-the-art, especially when the field is progressing at a rapid pace leaving many outside the cutting edge with an out-dated view of what contemporary capabilities are (and are not). This is a constructive and interdisciplinary effort, as the tools to measure performance of AI technologies are just as much in flux as the technology itself. Measurements of the technology focus either on performance (Eckersley & Nasser, 2017) or the resource use of the technology in terms of data or compute (Amodei & Hernandez, 2018), though other dimensions could also be measured (Martínez-Plumed et al, 2018). Other efforts go beyond the technology itself and also track the ecosystem in which the technology is developed, looking at hardware, conference attendance numbers, publications, enrolment, etc. (Benaich & Hogarth, 2018; Shoham et al, 2017).

2.5 Interactive futures narratives and scenarios

For most of the futures exploration tools described above, the audience is passive, and is being communicated at via text or vision and sound. Even surveys of the public often involve only localised and limited contributions from each individual. However, there also exist tools that enable the audience to take a more active role, either in a pre-defined narrative or in the co-creation of narratives. The emphasis on greater public participation is a key tenant of responsible research and innovation (Owen, Macnaghten, & Stilgoe, 2012) and it applies with force to the field of artificial intelligence (Stilgoe, 2018).

2.5.1 Participatory futures workshops

On the more formal end, participatory future workshops (Jungk & Müllert, 1987), or one of the numerous variations on the theme (Nikolova, 2014; Oliverio, 2017), go through a structured engagement between different stakeholders. These reflect the (originally more corporate and less open) processes of scenario planning (Amer, Daim, & Jetter, 2013). Similar to scenario planning, where participants explore a range of possible futures as a team, wargaming (Perla, 1990) and drama theory (Bryant, 2002) use role-play to place participants in opposing roles, to explore what strategies may emerge or investigate novel opportunities for cooperation and resolution. While the author knows of no such exercises on long-term AI futures, nearer-term exercises, for example on autonomous driving, are already taking place (Cohen, Stilgoe, & Cavoli, 2018). When such exercises have the support of government and buy-in from both experts and non-experts, they can prove to be highly valuable tools in preparing for AI futures; indeed, they come close to certain visions of the ideal interaction between

science and society (Kitcher, 2011). However, they also require significant resources and expertise to carry out well.

2.5.2 Interactive fictions

At the less participatory end, but still allowing the audience to play a more active role, are interactive fictions, especially in the medium of video games. While artificial intelligence, as a long-standing science fiction trope, has been depicted in video games for decades, recent games incorporate more of the nuanced arguments presented about the potential futures and characteristics of AI.

For example, *The Red Strings Club* (Deconstructoid, 2018) explores fundamental questions of machine ethics in an interactive dialog with the player, and *Universal Paperclips* (Lantz, 2017) allows the player to experience a thought experiment created to explore the "orthogonality thesis", the argument that arbitrarily high levels of intelligence are compatible with a wide range of ultimate goals, including ones that would seem to us foolish or nonsensical (Bostrom, 2014).

Other video games focus less on the narrative element, but rather present a rich simulator in which artificial intelligence is one of many technologies available to the player, allowing the exploration of a wide range of future AI scenarios and their interplay with other systems such as diplomacy or resource management. Examples include *Stellaris* (Paradox Interactive, 2016), in which artificial intelligence technologies are available to the player as they establish their galactic empire, or the *Superintelligence* mod (Shapira & Avin, 2017) for *Sid Meier's Civilisation V* (Firaxis Games, 2010), which allows the player, in the shoes of a world leader, to gain a strategic advantage using AI and achieve a scientific victory by creating an artificial superintelligence, while risking the creation of an unsafe superintelligence which can lead

to an existential catastrophe.

2.5.3 Role-play scenarios

While video games allow audiences to take a more active role in the exploration of possible AI futures within the game environment, they hardly satisfy the call for public participation in jointly imagining and constructing the future of emerging technologies. To explore AI futures in a collaborative and inclusive manner, experts and audiences must explore them together. One way to achieve this is through the joint exploration of narratives in role-play games.

Scenarios that have been developed with expert participation through any of the methods above, or through other means, can be circulated more broadly as templates for role-play games amongst interested parties. At the hobbyist level, game systems such as *Revolt of the Machines* (Fantasy Flight Games, 2016) and *Mutant: Mechatron* (Ligan, 2017) allow players to collectively explore a possible AI future. While these are often very entertaining, they may fall into the same failures as narrative fictions. It seems there is currently an unmet need for realistic and engaging AI futures role-play game systems.

3. Summary and conclusion

As AI hype drives utopian and dystopian visions, while rapid technological progress and adoption leaves many of us uncertain about the future impacts on our lives, the need for rich, informative, and grounded AI futures narratives is clear. It is also clear that there is a wide range of tools to develop such narratives, many of which are

available to creators and experts outside the AI research community. It is less clear, however, how best to utilise each of the available tools, with what urgency and in which domains. The table below summarises the different tools surveyed above, with their respective advantages and limitations.

Tool	**Existing abundance**	**Skills and resources required**	**Advantages**	**Limitations**
Fictional narratives	Overly abundant	Creative writing, production costs for film	Unbridled imagination, (relatively) open participation	Lack of realism, pull to extremes, lack of accountability, lack of diversity, skewed popularity distribution
Single-discipline futures exploration	Growing rapidly, though some disciplines are still missing	Domain expertise, familiarity with AI, forecasting skills	Deep dives into relevant facts and arguments	Predictive power is poor, disagreements can paralyse, not easy to integrate across disciplines
Surveys	Few key studies	Survey design, resources to carry out the survey	Aggregate evidence can counteract some biases, present a snapshot of current beliefs	Survey design is hard, topic in flux, misunderstanding is commonplace; poor predictive power
Interdisciplinary futures exploration	Few but growing rapidly	Interdisciplinary facilitation, network of stakeholders, time and geographic availability	Holistic view of complex topics, opportunity to directly engage with policy makers and other key stakeholders	Risk of groupthink, conservatism; scoping is difficult: too narrow and miss opportunities and challenges, too broad and becomes intractable
Evidence synthesis	Few	Access to studies in a range of disciplines and expertise to assess them and communicate findings	Evidence-based holistic picture drawing on a wide range of works, prepared with policy in mind	Time and labour intensive, evidence may be partial and rapidly changing, best practices still evolving

Extrapolating data trends	Few key hubs, abundant but disperse data	Familiarity with the field and the techniques of AI, measurement platforms, data harvesting and curation	Historical and contemporary measurements can be largely uncontested, informative	Difficult to extrapolate from past trends due to non-linearity, feedback, potential for discontinuity; need to constantly evolve and adapt measurements
Participatory futures workshops	None on long-term AI, few on short term issues such as self-driving cars	Buy-in from experts and non-expert participants, budget for workshops, facilitation skills, time of participants	Participatory, expert-informed exploration of future scenarios, legitimacy for policy guidance	Difficult to get buy-in and time commitment from experts and stakeholders, requires significant investment to tutor non-experts
Interactive fictions	Several, though few with realistic representations informed by recent advances	Game development skills and budget	Audience takes an active role, can explore alternatives, simulators offer a combinatorial explosion of options	Similar to fictional narratives, plus limitations of what can be represented effectively with limited skills and budget
Role-play scenarios	Few	Facilitation, game/scenario design	Stakeholders can come together to co-explore possible futures	Information gaps in the group can slow down or derail the conversation, strongly depends on the available expertise and facilitation skills

As can be expected, no tool is strictly better than all other tools. Some provide more evidence-based, deep analysis, but tend to be limited in the range of questions they can cover and place barriers on participation. Others allow for more diverse and integrative perspectives, but tend to preclude detailed and in-depth analysis or come at a very high cost in terms of time and facilitation. Instead of judging individual

futures narratives in isolation, it may be more useful to look at the entire ecosystem of future AI narratives, asking whether certain narratives are dominating our imagination without sufficient warrant, or if there are tools and narratives that are underutilised or gaining insufficient attention. At present, it seems that not enough attention is being given to data-driven, realistic, integrative, and participatory scenario role-plays, which can build on and integrate a range of other tools and narratives and make them more accessible to a wider audience in a more nuanced way. A more balanced portfolio is called for.

As we act to critique and curate the ecosystem of AI futures, we should keep in mind the aims of these narratives: beyond entertainment and education, there are real ongoing processes of technological development and deployment that currently have, are likely to continue to have, significant social impacts. These processes are not isolated from the societies in which they take place, and the interactions between technology developers, policymakers, diverse stakeholders and numerous publics are mediated and shaped by the futures narratives each group has access to. Thus, AI futures narratives play a crucial role in making sure we arrive at futures of our own choosing, that reflect our values and preferences, that minimise frictions along the path, and that do not take us by surprise. Thus, critique and curation of AI futures is an integral part of the process of responsible development of artificial intelligence, a part in which humanities scholars have a significant role to play.

Acknowledgements. I would like to thank three anonymous reviewers, Haydn Belfield, Simon Beard, and the attendees of the 1st International Conference on AI Humanities for helpful comments.

References

Achen, C. H., & Bartels, L. M. (2017). *Democracy for realists: Why elections do not produce responsive government* (Vol. 4). Princeton University Press.

Amer, M., Daim, T. U., & Jetter, A. (2013). A review of scenario planning. *Futures*, *46*, 23-40.

Amodei, D. & Hernandez, D. (2018). AI and Compute. Open AI blog. Retrieved from https://blog.openai.com/ai-and-compute/

Armstrong, S., & Sotala, K. (2015). How we're predicting AI–or failing to. In *Beyond Artificial Intelligence* (pp. 11-29). Springer, Cham.

Baum, S. (2018). Superintelligence skepticism as a political tool. *Information*, *9*(9), 209.

Barrett, A. M. & Baum, S. D. (2017). A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, *29*(2), 397-414.

Bastani, A. (2018). *Fully Automated Luxury Communism: A Manifesto*. Verso.

Benaich, N. & Hogarth, I. (2018) The state of artificial intelligence in 2018: A good old fashioned report. Retrieved from https://www.stateof.ai/

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Boden, M. A. (2016). *AI: Its nature and future*. Oxford University Press.

Bostrom, N. (2014). Superintelligence: paths, dangers, strategies.

British Academy and The Royal Society (2018) The impact of artificial intelligence on work. Retrieved from https://royalsociety.org/~/media/policy/projects/ai-and-work/evidence-synthesis-the-impact-of-AI-on-work.PDF

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Anderson, H. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint*

arXiv:1802.07228.

Bryant, J. (2002). *The six dilemmas of collaboration: Inter-organisational relationships as drama*. Wiley.

Campbell, M., Hoane Jr, A. J., & Hsu, F. H. (2002). Deep blue. *Artificial intelligence*, *134*(1-2), 57-83.

Cave, S. & Dihal, K. (2018). Ancient dreams of intelligent machines: 3,000 years of robots. *Nature*, *559*(7715), 473.

Cohen, T., Stilgoe, J., & Cavoli, C. (2018). Reframing the governance of automotive automation: insights from UK stakeholder workshops. *Journal of Responsible Innovation*, 1-23.

Collins, H. M. & Evans, R. (2002). The third wave of science studies: Studies of expertise and experience. *Social studies of science*, *32*(2), 235-296.

De Vany, A. (2004). *Hollywood economics: How extreme uncertainty shapes the film industry*. Routledge.

Deconstructeam (2018) The red strings club [PC game]. Devolver Digital.

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

Donnelly, C. A., Boyd, I., Campbell, P., Craig, C., Vallance, P., Walport, M., ... & Wormald, C. (2018). Four principles to make evidence synthesis more useful for policy. *Nature*, *558*(7710), 361.

Eckersley, P. & Nasser, Y. et al. (2017). EFF AI progress measurement project. Retrieved from https://eff.org/ai/metrics

Endy, D. (2014). Synthetic biology – What should we be vibrating about?: Drew Endy at TEDxStanford, Retreived from https://www.youtube.com/watch?v=rf5tTe_i7aA.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115.

Fantasy Flight Games. (2016). End of the world: Revolt of the machines [roleplaying game].

Farmer, J. D., & Lafond, F. (2016). How predictable is technological progress?. *Research Policy*, *45*(3), 647-665.

Felten, E. & Lyons, T. (2016). The administration's report on the future of artificial intelligence. Retrieved from

https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence

Firaxis Games (2010) Sid Meier's Civilization V [PC game].

Fraze, D. (2018). Cyber grand challenge (CGC). Retrieved from https://www.darpa.mil/program/cyber-grand-challenge

Fry, H. (2018). *Hello World: How to be Human in the Age of the Machine.* Penguin.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When will AI exceed human performance? Evidence from AI experts. *arXiv preprint arXiv:1705.08807*.

Grosz, B. J., & Stone, P. (2018). A Century Long Commitment to Assessing Artificial Intelligence and its Impact on Society. *arXiv preprint arXiv:1808.07899*.

Harris, J. (2017). 16 AI bots with human names. Chatbots Life. Retrieved from https://chatbotslife.com/10-ai-bots-with-human-names-7efd7047be34

Hanson, R. (2016). *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press.

House of Lords. (2018). AI in the UK: ready, willing and able? House of Lords Select Committee on Artificial Intelligence Report of Session 2017–19

Hyacinth, B. T. (2017). *The Future of Leadership: Rise of Automation, Robotics and Artificial Intelligence*. MBA Caribbean Organisation.

Jungk, R., & Müllert, N. (1987). *Future Workshops: How to create desirable futures*. London: Institute for Social Inventions.

Kakoudaki, D. (2014). *Anatomy of a robot: Literature, cinema, and the cultural work of artificial people*. Rutgers University Press.

Kareiva, P., & Carranza, V. (2018). Existential risk due to ecosystem collapse: Nature strikes back. *Futures*.

Kirby, D. A. (2011). *Lab coats in Hollywood: Science, scientists, and cinema*. MIT Press.

Kitcher, P. (2011). *Science in a democratic society*. Prometheus Books.

Kurzweil, R. (2010). *The singularity is near*. Gerald Duckworth & Co.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Lantz, F. (2017) Universal Paperclips [online video game]. Retrieved from http://www.decisionproblem.com/paperclips/

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*(4), 391-444.

Ligan, F. (2017). Mutant: Mechatron [roleplaying game].

Lippmann, R. (1987). An introduction to computing with neural nets. *IEEE Assp magazine*, *4*(2), 4-22.

Lyrebird (2018) We create the most realistic artificial voices in the world. Retrieved from https://lyrebird.ai/

Martínez-Plumed, F., Avin, S., Brundage, M., Dafoe, A., hÉigeartaigh, S. Ó., & Hernández-Orallo, J. (2018). Accounting for the Neglected Dimensions of AI Progress. *arXiv preprint arXiv:1806.00610*.

Minsky, M. (1988). *Society of mind*. Simon and Schuster.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Mozur, P. (2018, July 8). Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras. *New York Times.*

Müller, V. C. & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence* (pp. 555-572). Springer, Cham.

Needham, D. J. & Weitzdörfer J. F. W., eds. (forthcoming) *Extremes.* Cambridge University Press.

Nikolova, B. (2014). The rise and promise of participatory foresight. *European Journal of Futures Research*, *2*(1), 33.

Oleszczuk, A. (2017). Sad and Rabid Puppies: Politicization of the Hugo Award Nomination Procedure. *New Horizons in English Studies*, (2), 127.

Oliverio, V. (2017) Participatory Foresight. Centre for Strategic Futures. Retrieved from https://www.csf.gov.sg/our-work/Publications/Publication/Index/participatory-foresight

Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and public policy*, *39*(6), 751-760.

Owens, S. (2011). Three thoughts on the Third Wave. *Critical policy studies*, *5*(3), 329-333.

Paradox Interactive (2016) Stellaris [video game].

Perla, P. (1990). *The art of wargaming: a guide for professionals and hobbyists* (Vol. 89028818). Annapolis, MD: Naval Institute Press.

Rose, H. (2000). Science Fiction's memory of the future. In *Contested Futures: A sociology of prospective techno-science. Ashgate, Aldershot, UK*, 157-174.

Rouhianien, L (2018) *Artificial Intelligence: 101 Things You Must Know Today about Our Future.* Createspace Independent Publishing Platform.

Rowe, T. & Beard, S. (2018). Probabilities, methodologies and the evidence base in existential risk assessments. Working Paper. Retrieved from http://eprints.lse.ac.uk/89506/

Shanahan, M. (2010). *Embodiment and the inner life: Cognition and Consciousness in the Space of Possible Minds*. Oxford University Press, USA.

Shanahan, M. (2015). *The technological singularity*. MIT Press.

Shapira, S. & Avin, S. (2017) Superintelligence [video game mod]. Retrieved from https://steamcommunity.com/sharedfiles/filedetails/?id=1215263272

Shoham, Y., Perrault, R., Brynjolfsson, E., & Clark, J. (2017). Artificial Intelligence Index－2017 Annual Report. Retrieved from https://aiindex.org/2017-report.pdf

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354.

Srnicek, N. & Williams, A. (2015). *Inventing the future: Postcapitalism and a world without work*. Verso Books.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social studies of science*, *48*(1), 25-56.

Sutherland, W. J., & Wordley, C. F. (2018). A fresh approach to evidence synthesis. *Nature 558*, 364-366

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, *36*(4), 95.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence.*

Knopf.

The Royal Society (2017) Public views of machine learning. Retrieved from https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

Turchin, A. & Denkenberger, D. (2018). Classification of global catastrophic risks connected with artificial intelligence. *AI & SOCIETY*, 1-17.

Turing, A.M. (1950) Computing Machinery and Intelligence. *Mind* 49: pp.433-460.

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks*, *1*(303), 184.

Yudkowsky, E. (2017). There's no fire alarm for artificial general intelligence. *Machine Intelligence Research Institute*. Retrieved from https://intelligence.org/2017/10/13/fire-alarm/

Zhang, C., Li, H., Wang, X., & Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 833-841).