European Parliament

Why artificial intelligence matters

The term AI refers to a variety of methods, applying equally well to self-replicating robots that could have extreme implications for humanity centuries from now as it does to tools for early diagnosis of cancer. Both are called AI, but the similarities end there. They deploy different methods in different contexts with different aims and different consequences. Most notably the former is a speculated and unlikely future, while the latter is a fact – it is already here. All kinds of AI are worthy of discussion but, for a meaningful and constructive debate, arguments should be grounded in the specific methods, contexts, timeframes and probabilities that they concern.

This briefing explains why AI matters by reviewing some of the key opportunities and challenges it presents, but it does so with reference to the functionality and readiness of the technology. The first section focuses on the opportunities and challenges presented by today's AI while the second explores longer-term speculative opportunities and challenges that are contingent upon future developments that may never happen. Readers that are unfamiliar with the state of the art can consult the companion briefing, How artificial intelligence works.

Current opportunities and challenges

The primary reason why AI matters is because of the immense potential it presents, both currently and speculatively, to benefit our lives. This includes serious benefits – such as supporting more effective health, production, transportation and decision-making systems – as well as more frivolous benefits such as minor efficiency gains and novelty or entertainment value provided by a proliferation of 'gadgets'. Nonetheless, even apparently inane examples can also provide indirect benefits by generating capital, expertise and data which can contribute to more serious application areas in the future. For example, tools trained to identify images of zebras and dishwashers can be redeployed to identify cancers, and AI capacity gained through developing game playing AI can be redeployed in healthcare. Of course, the same disruptions also present legal, social, ethical and economic challenges. These are sometimes related to the technology itself, with questions of transparency, bias and autonomy, or to the business models, which often prioritise gathering data or targeting advertising rather than delivering genuine social value.

It is important to note that, even if a great deal of time is spent discussing the challenges, this effort is justified by the widely recognised opportunities presented by the technology. Simply put, if AI development was only bad news, then it would not be developed and the challenges would not be worth discussing. In recognition of AI's benefits as the *raison d'être* for the debate about AI, the first challenge presented here is to avoid unnecessary underuse of AI.

Countering unnecessary underuse

Discussions of AI – at least in policy contexts – are often dominated by challenges. This can give a superficial impression of negativity, or at least defensiveness. However, the core motivation of these debates is always to maximise the benefits of the technology to society. This does not mean supporting AI at all costs, but balancing its positive and negative impacts while taking account of a wide range of uncertain social, economic and technical trends. Naturally, there are different opinions on what constitutes a cost or a benefit, as well as how best to distribute them, but any

Author: Philip Boucher
Scientific Foresight Unit (STOA)
PE 634.421 – March 2019

unnecessarily unexploited benefits could quickly add up to substantial opportunity costs. This underuse could be inadvertent or deliberate.

Inadvertent underuse could result from failures to make the right strategic choices with regards AI, or to features of the wider cultural, economic, technical or political context that do not lend themselves to AI development. In a European context, this could be due to the difficulties presented by fragmented internal markets, which could limit economic and technical competitiveness with larger markets such as those in the USA and China. Take the example of using national medical records to develop health screening AI. The USA and China have more records than EU Member States, and so have a technical advantage. Indeed, while Europe does maintain an important role in global AI development, particularly in terms of fundamental research, it is widely recognised that the USA and China dominate the frontline of global AI development. This is often explained with reference to their higher levels of investment, lower levels of data protection, and an appetite for application and rapid adoption.

Deliberate underuse of AI differs in that maximum use of AI is strategically avoided. This could be motivated by a desire to prioritise alignment with certain principles or values, or prompted by fears that may or may not turn out to have been legitimate. Continuing the example of national health records, Member States may prefer to impose restrictions on the transfer and processing of sensitive medical data. This may have short-term costs in terms of restricting market activity and slowing technical advancement. However, these costs may be worth bearing for ethical reasons, and may be recouped in the long-term, for example through greater consumer trust and confidence in the tools that are eventually developed. Other forms of deliberate underuse may include moratoriums on more controversial applications of AI, for example in the development of autonomous weapons or artificial consciousness (both of which are further discussed below).

Transparency, explainability and responsibility

Today's AI presents a range of different transparency challenges. Perhaps the most salient is the lack of explainability of AI, that is, how the internal decision-making logic of an AI agent can be understood and described in human terms. This challenge is a function of AI methods. For example, artificial neural networks process images pixel-by-pixel, performing millions of calculations before deciding whether or not it depicts a dog. While the AI is often right, it is very difficult – even for the engineers that designed the system – to translate their internal logic into an argument that makes sense to humans. In many cases, it can be impossible to translate millions of calculations into a an explanation fit for human experts, let alone for users, policy-makers, judges and juries. Since the current generation of machine learning AI are not generally able to explain their internal logic in a human-friendly format, their role in decision-making remains opaque.

While a technical lack of transparency may be an unfortunate side effect of AI methods, there are also many deliberate and forms of opacity in AI decision-making. This second transparency challenge is more strongly related to commercial interests and business practices. Since AI enables large-scale automated categorisation, it can be used to treat individuals differently. For example, customers 'willingness to pay' for items can be estimated through analysis of their shopping practices and other indicators. By sending individual discount coupons to their customers, shops can effectively establish individualised pricing regimes that reduce prices to their clients' estimated willingness to pay. Similar trends are seen in political campaigns, as citizens are presented with customised messages based upon AI estimations of what would convince them to vote for a specific candidate. Where different promises are made to different citizens, it is difficult to hold politicians to account. Furthermore, campaigns can be run by third parties such as foreign powers or commercial interests, outside the control of the politician, to influence the outcome of elections. In both examples, the customer and citizen are deliberately restricted from accessing all available information, be it about prices or promises, in order to gain a commercial or strategic advantage.

A third transparency challenge is identified in individuals not always knowing whether they are interacting with an AI or human agent. This can include chat interfaces, but can also extend to other

forms of interaction, such as the processing of loan or job applications. Where clients are aware that they are interacting with an AI agent, they may still be restricted from understanding how and why the decisions are made, either because it is too complicated (as in the first transparency challenge) or because access to its internal functioning is restricted by legal protection for intellectual property (related to the second transparency challenge).

Finally, there are also longer-term transparency challenges related to strategic opacity about the full range of intended and expected outcomes of AI development. Following the principles of responsible research and innovation, meaningful public debate requires transparency about the full range of expected outcomes of development and application. While it can be tempting to encourage public acceptance by focusing on the benefits of AI, the approach is neither responsible nor effective. To earn trust and achieve informed consent, it is important to communicate potential drawbacks. This is related to the next challenge, that of the public acceptability of AI.

Public opposition and acceptability

Public opposition is regularly cited as a current challenge for AI, often explained in terms of lack of understanding of the technology and appreciation for its benefits. This 'deficit model' is a common sentiment in technology development, and has been extensively studied and repeatedly criticised as inaccurate and ineffective. Inaccurate because public opposition to technologies is not usually characterised by misunderstanding but a lack of control over outcomes, and ineffective because repeating positive messages without recognising problems can lead to more entrenched positions. Where the messages are imbalanced – for example by avoiding discussion of expected outcomes that are anticipated to raise objections – the approach can be characterised as deliberately misleading and can contribute to further public mistrust.

It is worth mentioning that AI does not currently face substantial public opposition. Critical voices usually come from experts and stakeholders, and their objections tend to either be speculative, or focused on wider issues – such as citizens' control of data, the distribution of costs and benefits, the concentration of power, or military activities – rather than the technology per se. It is possible that AI is not subject to significant public opposition because people accept what they understand to be the costs, benefits and uncertainties. On the other hand, since AI is diffused through personal devices, commercial infrastructures and public services, it may be less susceptible to the kind of action that is mobilised against discrete, localised and physical technologies, such as GM crops and nuclear power stations. People that are uneasy about AI may feel powerless to opt out, or to shape the role of AI in their lives in any meaningful way.

Algorithms can reinforce bias and inequality

Generally speaking, AI engineers do not deliberately produce prejudiced algorithms. As explained in the companion briefing, algorithms learn how to make decisions by following patterns in the data that is provided to them. As such, it is unsurprising that when previous patters of bias and inequality can be detected in data then the resulting algorithms can also be biased. In an illustrative example, Microsoft released Tay, an AI bot that learnt to chat by analysing and engaging in conversations with humans on Twitter. Within 24 hours, Tay spoke like an angry, confused, racist misogynist. Unfortunately, there are many biases and inequalities in our societies, and the unpleasant chatbot reflects the unpleasant human interactions on the internet that formed its training environment. Likewise, an AI trained to identify promising engineers or potential criminals will reflect the structural inequalities – based on gender, ethnic origin, age and other factors – that are evident in data about engineers and criminals.

The chatbot was quickly switched off, but there are many well-documented examples of algorithms that exacerbate existing biases and reinforce inequalities, often disproportionately affecting the most marginalised members of society. In an illustrative examples from the book *Weapons of math destruction*, an algorithm for identifying how likely a prisoner is to reoffend was introduced to support more objective parole decisions, but was shown to disproportionately discriminate against

black inmates. The case showed how, even where information about race is not directly available to the algorithm, proxies from other details such name, address and school can be enough to train a biased algorithm. Indeed, several studies have shown that it is possible to de-anonymise data or make accurate predictions about user's private lives with reference to very few variables. As discussed in the section on transparency, the internal logic of these systems are often not open for examination – either because of their technical complexity or for commercial reasons – which limits the potential for redress.

Algorithms are not objective because, just like people, in the course of their training they develop a way of making sense of what they have seen before, and use this 'worldview' to categorise new situations with which they are presented. The fact that algorithms have subjective worldviews is often underappreciated. They might appear objective because they apply their bias more consistently than humans. Perhaps their use of numbers to represent complex social reality gives them an air of precise facticity and – since humans recognise their impressive power but find it difficult to comprehend their logic – they simply yield to their superiority. Understanding that AI agents are inherently subjective is a crucial prerequisite for ensuring that they are only applied to tasks for which they are well equipped. After all, if AI are discrimination machines, we prefer to set them to discriminate against cancer rather than vulnerable people.

Informed consent: Privacy and human experimentation

The use of data about individuals in AI algorithms raises several challenges related to citizens' giving their informed consent for their data to be stored, processed or shared for particular purposes. Informed consent is also related to the transparency challenge, as truly informed consent requires that the individual is aware of and understands the situation to which they consent. Taking this concept seriously requires that the full range of expected outcomes is communicated to the individual in a way they can understand. Since AI tools are often data-oriented (both in their development and application), one of the expected outcomes of their use is the processing of personal data, or information pertaining to the user's private life, such as their credit rating, medical records or relationships. Finding meaningful ways of gaining informed consent for this use can be problematic, as illustrated by the routine acceptance of cookies and agreement to terms and conditions which are not well understood in exchange for access to information or services.

Another aspect of the informed consent challenge concerns the use of AI for research purposes, whereby the data subject may also be considered a research subject. Following best practice in medical research, participants in experiments are protected by principles of informed consent. So, for example, when Facebook conducted an experiment on its users – presenting some with positive content and others with negative content – to analyse whether they could control the emotions of its users (result: they could), they breached the ethical principle of conducting psychological experiments on people without their knowledge or permission. Furthermore, such experiments can be difficult to identify unless they are disclosed voluntarily.

Military applications and security issues

AI is a dual use technology, so advances in civilian AI will help develop military AI, just as advances in military AI will help develop civilian AI. These synergies have existed since the earliest days of AI. Historically, civilian AI developments mostly followed military development. Indeed, following the protagonists of AI history – from Florence Nightingale to Alan Turing – many of the applied mathematics techniques that form the basis of contemporary AI were developed in the context of war. At present, the reverse of this relationship is more prevalent, as civilian AI is adapted and applied to military applications. Taking the example of drone guidance systems, the same techniques designed to autonomously 'sense and avoid' in-air collisions can also be deployed for the autonomous control of target acquisition.

Whatever the mechanisms of dual use synergies, AI has an increasing role in cybersecurity, information warfare and physical combat. In cybersecurity, AI can play an important role in both the

attack and defence strategies of hacking, phishing and other types of security breach. Information warfare can also make use of AI, for example in the use of bots to influence public discourse or behavioural profiling for targeting political messages, as seen in the Cambridge Analytical scandal. This is a particularly difficult challenge for liberal Western democracies to counter, as methods of responding to hostile information campaigns can be accused of curtailing freedom of speech.

AI is not only important in the digital domains of security and warfare, but also increasingly in physical combat. This does not mean robotic humanoid soldiers – a speculative vision that has no basis in current AI capabilities – but rather to the use of AI today in, for example, autonomous systems for guiding rockets or drones. These systems vary in their level of autonomy. In 'human in the loop' systems, key decisions such as firing weapons are always made by a person, whereas 'human on the loop' systems operate autonomously under the supervision of a human that can interrupt activities, and 'fire and forget' systems complete their mission without any human supervision. The latter are already in use for weapons systems, for example the Harpy drone which seeks and destroys radar systems without human control or supervision.

The civilian drone market is highlighted as the key source of both innovation and financing for future military technology developments. This approach can be seen across the spectrum of military AI. DARPA – the US military research agency – funds activities to stimulate specific civilian AI developments with military AI as the ultimate beneficiary, and has a US$2 billion investment strategy to embed AI in weapons systems. Similarly, the US Joint Artificial Intelligence Center is tasked with accelerating the application of AI research and development across the US military.

Some may consider military AI – such as drones that can strike targets autonomously without human approval – to cross red lines of public acceptability, ethics or military rules of engagement. This issue is revisited in the context of speculative challenges, but even where the dual use status of AI technology is not considered a problem in itself, if it is deliberately downplayed as part of a public acceptance strategy, this could violate principles of responsible innovation, which require that the full range of expected outcomes of development are clearly communicated so that an informed debate can take place.

Incumbent's advantage

Today's AI development is largely driven by data. This data is usually gained by a business model whereby free access to applications is granted to users in exchange for their data and exposure to advertisements. The more widely used a service, the more data it can gather and deploy, which in turn enables new services to be developed and offered to both users and advertisers. These services attract more users, and the cycle of data collection and application continues. This dynamic favours bigger players in the market, a situation that is exacerbated by the global character of such firms, which enables them to develop more tax efficient strategies than localised firms.

The same mechanisms are in play for AI in the public sector as, for example, AI health applications require medical data, but these are strictly regulated and not readily shared between public services, let alone with private companies or other nations. As such, larger health services have more 'in-house' data and, all else being equal, can develop better AI tools. Returning to private sector AI, the feedback loop for social networks such as Facebook are particularly favourable to the incumbents, as the large membership itself attracts more members, and increases the cost to customers who want to change network. This can lead to a rather extreme concentration of resources, whereby the owners of the data have access to substantial information about users, significant control over the information that they receive and the choices they have, and even the capability to 'nudge' their emotional states. On the other hand, users have limited access to, or control over, this data and its application and competitors are not in a position to offer rival services to which users can switch without losing access to the contacts and data that are mediated by the incumbents.

Countering damaging overuse

Knowledge of how AI works can have a demystifying effect. Realising extreme utopian or dystopian AI fantasies would require not just incremental improvements, but immense paradigm shifts in how AI works. This is mostly reassuring, but also introduces a new set of challenges. Once the smart mathematical 'tricks' of AI are revealed, their limitations come into focus. Many people may overestimate the capabilities of today's AI to understand the world and make sound decisions. Indeed, AI barely lives up to the loosest definitions of intelligence. What if society has too much confidence in AI, over-relying on it in some areas and introducing it to others where it is simply not up to the job? While the first challenge discussed here was to counter the unnecessary underuse of today's AI, the final challenge is to counter its damaging overuse.

One form of overuse in AI is due to overconfidence in its capabilities. Where AI agents (and AI powered robots) are designed to mimic humans, for example by speaking or moving in ways that correspond to human expressions of emotion, it can be tempting to imagine that the AI is also 'feeling' the emotions that correspond to the cues that they give. This can provoke empathy, trust and other strong forms of meaningful engagement from humans that the agent is simply incapable of reciprocating. Similarly, AI agents can process words very effectively, and even produce interesting images and develop novel strategies. However, it is important to recognise that even though the AI can perform these tasks at a high level, they do not comprehend the words, images and ideas in the same ways that humans do. The impressive feats of AI can inspire overconfidence in its capabilities, which can lead to its application in tasks for which is not well suited.

A second form of overuse concerns tasks for which AI is well suited, but its widespread application introduces vulnerabilities. A well-known example is the AF447 air crash in 2009, which was partially blamed upon the over-reliance of the human pilots on automated flight control systems. The pilots relied heavily on the automated system and, when it malfunctioned, they were unable to take over seamlessly, leading to the crash and 228 fatalities. Where reliable AI takes over tasks, in the long-term, humans have less chance to develop experience and also tend to increasingly trust the AI over their own judgement. This creates vulnerabilities in emergency situations where the AI malfunctions, but also in normal situations, as authority is tacitly ceded to the machine because it is assumed to make better decisions. Whether this overuse constitutes a slow 'surrender' to AI that is insufficiently capable brings us towards the realm of speculative challenges and opportunities.

Speculative challenges and opportunities

Today's AI agents can do many useful things with images and language, and perform many tasks at a very high level, but they cannot understand or create in the same ways as humans, and have trouble contextualising problems and explaining themselves in a human-friendly way. Speculative AI tends to assume that AI will be developed that can truly do these things, and often suggests that they will be effectively combined with precise and robust robots, which are equally speculative. Even incremental improvements within the current AI paradigm would likely fall short, because they require paradigm-shifting development. The companion briefing, How artificial intelligence works, reviews some potential avenues for such advancement, including self-explaining, context-sensitive, robotic and quantum AI. With these speculated powers come speculated opportunities and challenges. This section provides a review of the most salient examples, which are usually extrapolated (and often extreme) versions of those presented by today's AI.

Why are future scenarios so dystopian or utopian?

Future AI are often discussed in the context of a dystopian runaway AI that is capable of improving itself and escaping human control, with disastrous consequences. One example is systems with advanced minds and forms of consciousness that can develop their own goals, which might not align with those of their human creators. Perhaps this AI could anticipate human concerns and the possibility of being switched off, and hide the true extent of their capabilities until they are too

powerful to be controlled. Another example speculates AIs that stick to the objectives set by humans, but do so with a level of capability and autonomy that innocent goals such as 'create paperclips' could lead perverse outcomes, such enslaving humanity in paperclip factories or even transforming all earthly matter into stationery. Either way, these scenarios present the power and autonomy of the AI as an existential threat to humanity.

Future AI scenarios are often criticised, particularly by representatives of the technical and commercial sectors, as being too pessimistic and dystopian. However, these sectors are often the source of extremely optimistic and utopian speculations about future AI. These range from suggesting that AI can make undesirable labour obsolete, freeing humans to focus on whatever they want, notably creative, scholarly, and leisurely pursuits; as well as tasks where human contact and interaction is considered essential, often in relation to children, the elderly and those in need of social, physical or medical support. To a large extent, all long-term AI scenarios only really work because people find it so easy to imagine substantial medium-term advancement of AI, which takes an increasingly important role in all aspects of our lives. In this sense, our belief in the immense capability and utility of AI is a prerequisite for even the bleakest dystopian visions.

Runaway AI scenarios make good films and snappy headlines, but there remains an immense gap in the capability and autonomy of the paperclip monstrosity, for example, and the state of the art. However, the same gap also exists between today's AI and those that can be seamlessly embedded in our societies, making us live longer, happier, wealthier and healthier lives. If the former is an unrealistic scare story, the latter is an unrealistic marketing story. Collins' book, *Artifictional Intelligence*, laments the lack of epistemic modesty demonstrated in the AI community. Rather than highlighting what is not understood and what needs to be improved, the field basks in its limited achievements and makes promises for society that are beyond its capabilities. If, Collins argues, AI was more modest, more demanding of itself, then it could perhaps achieve more and be less vulnerable to damaging overuse.

There are several answers to the question of why future AI scenarios are so dystopian or utopian. First of all, this is not really the case, but the more measured and realistic visions are too mundane to appear in films, headlines or advertisements, so they generate less hype and attract less attention. Second, even if there is only a very small chance of them occurring, some potential impacts are so serious that they demand at least some reflection. This is a basic tenant of risk assessment and it applies to both pessimistic visions, such as loss of human autonomy, and optimistic visions, such as 'workless society'. Finally, the effort of creating and reflecting on visions that will never come to pass is not wasted. The scenarios can help us to respond to less extreme variants which are relevant to today's AI, such as increased decision-making by AI and changing work patterns. They can also help us to make sense of our societies' evolving relationship with technology, and might even help us to identify desirable futures that we could work towards.

Making employment obsolete

Disruption to employment models is a crucial aspect of many future AI scenarios. In some visions, human workers are replaced by AI agents that do not take holidays, join unions or even draw salaries. The scenario leads to more unequal societies, as those with a stake in the means of production grow wealthy, while displaced workers face unemployment and poverty. Worse, an irrelevant underclass could emerge, which loses its negotiating position along with its role in the production system. For the optimists, job losses are not a problem if the very concept of employment is also made obsolete and all citizens can live off the basic productivity of AI. Some may choose to work because they like it, or desire more than the basics, while others could devote their lives to their relationships, artistic pursuits, leisurely and sporting activities. In fact, these two visions share much common ground, both revolving around the distribution of costs and benefits. Indeed, the two scenarios can be easily combined in a single vision, where some nations profit from AI development and use their resources to provide safety nets for their citizens while other countries fall behind, leading to pockets of extreme wealth and extreme poverty in different parts of the world.

Another common scenario for the impact of future AI on employment sees jobs changing, as they have continuously since the industrial revolution, rather than disappearing. The mantra that echoes around these discussions is that some jobs will be lost but other, more fulfilling and high quality jobs will be created. This scenario tends to refocus debate on how to develop flexible workforces that can adapt to changing workplace requirements. However, while this scenario may appear less extreme than the leisure society or unemployment versions, it lacks a serious empirical basis and is no less fanciful than other speculative AI futures.

Human autonomy and security

There are concerns that AI could develop so much capability and power in the future that humans could lose autonomy. The paperclip dystopia described above is an extreme form of this challenge, but there are also more moderate expressions which extend more realistically from one of the challenges of today's AI – avoiding damaging overuse. Advances in AI technology, combined with deeper integration in society, could lead to a dependence on AI that makes it difficult for humans to assert authority over decisions made by algorithms. Extending from another challenge of today's AI – the incumbent's advantage – future development of advanced AI to control transport, security or other major transversal systems would require consolidation of massive amounts of data and power. Current momentum suggests that such systems would likely be owned by large international companies, although they could conceivably be developed or placed under the control of individual nation states or international organisations.

The counterpoint to this challenge is the speculative opportunity based upon the idea that human autonomy and machine autonomy does not have to be a zero sum game. AI can be developed to enhance human autonomy by facilitating better decision-making. This may require development in self-explanatory and context sensitive AI, as well as new approaches to integrating digital tools into workplaces.

Crossing 'red lines'?

The final speculative challenge here is the possibility that AI could cross lines that are considered fundamentally unacceptable. Suggested red lines include the development of artificial consciousness which is capable of potentially infinite artificial suffering, as well as the deployment of fully autonomous weapons that deploy lethal force. The former is beyond current AI capabilities, although there is active research in the field, while the latter is certainly within the capabilities of today's AI and subject to substantial debate. Assuming widespread agreement on what and where these red lines are – and this is another speculation – the mechanisms of enforcing them would also present serious challenges. Development in controversial domains could be pushed underground, or relocated to nations that are unable or unwilling to enforce moratoria

stoa@ep.europa.eu (contact)
http://www.europarl.europa.eu/stoa/ (STOA website)
www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)