7/6/2021

Martin W

Responsible AI

TTA

Contents

[1. Responsible AI 2](#_Toc78140928)

[2. Find instances where AI has failed? Or been used maliciously or incorrectly. 3](#_Toc78140929)

[3. Implications of when AI fails. There is a specific article in the GDPR Law that covers this, especially with automated decision making. (opt in and out options). 3](#_Toc78140930)

[4. What should organisations do to ensure that they are being responsible with AI and the wider use of data in general 6](#_Toc78140931)

[References 7](#_Toc78140932)

# Responsible AI

When we talk about AI, we usually mean a machine learning model that is used within a system to automate something. AI can be used for anything from insurance underwriting to detecting cancer. The defining characteristic is that there is limited\no human input into the decisions made by the system. This can lead to many potential issues and companies need to define a clear approach to using AI. Responsible AI is a governance framework aimed at doing exactly that.

The framework can include details on what data can be collected and used, how models should be evaluated, and how to best deploy and monitor models. The framework can also define who is accountable for any negative outcomes of AI. Frameworks will differ between companies. Some will define specific approaches and others will be more open to interpretation. They all seek to achieve the same thing. That is to create AI systems that are interpretable, fair, safe and respectful of a user’s privacy.[[1]](#endnote-1)

Chart

Description automatically generated

Figure : Responsible Governance AI Framework

The first area mentioned is **interpretability**. When we interpret a model, we get an explanation for how it makes predictions. An AI system could reject your application for a mortgage or diagnose you with cancer. A user would likely demand an explanation even if these decisions are correct. Some models are easier to interpret than others making it easier to get explanations. Responsible AI can define how we build interpretable models or when it is okay to use one that is less interpretable.

Related to interpretability is model **fairness**. It is possible for AI systems to make decisions that discriminate against certain groups of people. This bias comes from bias in the data used to train models. In general, the more interpretable a model the easier it is to ensure fairness and correct any bias. We still need a Responsible AI framework to define how we evaluate fairness and what to do when a model is found to make unfair predictions. This is especially important when using less interpretable models.

**Safety** and security is another concern. These are not new to software development and are address by techniques like encryption and software tests. The difference is that, unlike general computer systems, AI systems are not deterministic. When faced with new scenarios they can make unexpected decisions. The systems can even be manipulated to make incorrect decisions. This is particularly concerning when we are dealing with robots. If they make errors, things like self-driving cars can cause injury or death.

The last aspect is **privacy** and data governance. The quality of data used is important. If there are mistakes in the data used by AI, then the system may make incorrect decisions. In general, AI should also not be allowed to use sensitive data (e.g. medical history, trade-union membership). In Europe, many of these concerns are address by [GDPR](https://gdpr.eu/). Outside of Europe, they will need to be addressed by a company’s own Responsible AI framework.

[491]

# Find instances where AI has failed? Or been used maliciously or incorrectly.

A leading facial-recognition  technology[[2]](#endnote-2) recognized three-time Super Bowl champion Duron Harmon of the New England Patriots, Boston Bruins forward Brad Marchand, and 25 other New England proficient athletes as criminals. Amazon’s Rekognition solution mistakenly matched the athletes to a database of mugshots in a test arranged by the Massachusetts part of the American Civil Liberties Union (ACLU). Almost one-in-six players were wrongly distinguished.

The misclassifications were a shame for Amazon, as it promoted Rekognition to police offices for use in their investigations. This technology is one such example of AI gone bad and was proved flawed and was not encouraged to be used by the government officials without protections.

# Implications of when AI fails. There is a specific article in the GDPR Law that covers this, especially with automated decision making. (opt in and out options).

Artificial intelligence (AI) is doing a lot of good and will continue to provide many benefits for our modern world, but along with the good, there will inevitably be negative consequences.

Legendary physicist Stephen Hawking[[3]](#endnote-3) shared this ominous warning: “Success in creating effective AI could be the biggest event in the history of our civilisation. Or the worst. So, we cannot know if we will be infinitely helped by AI or ignored by it and side-lined, or conceivably destroyed by it.”

The first step in being able to prepare for the negative impacts of artificial intelligence[[4]](#endnote-4) is to consider what some of those negative impacts might be. Here are some key ones:

#### **AI Bias**

Since AI algorithms are built by humans, they can have [built-in bias](https://searchenterpriseai.techtarget.com/definition/machine-learning-bias-algorithm-bias-or-AI-bias)[[5]](#endnote-5) by those who either intentionally or inadvertently introduce them into the algorithm. If AI algorithms are built with a bias or the data in the training sets, they are given to learn from is biased, they will produce results that are biased. This reality could lead to unintended consequences like the ones we have seen with discriminatory recruiting algorithms and Microsoft’s Twitter chatbot that became racist. As companies build AI algorithms, they need to be developed and trained responsibly.

#### **Loss of Certain Jobs**

While many jobs will be created by artificial intelligence[[6]](#endnote-6) and many people predict a net increase in jobs or at least anticipate the same amount will be created to replace[[7]](#endnote-7) the ones that are lost thanks to AI technology, there will be jobs people do today that machines will take over. This will require changes to training and education programmes to prepare our future workforce as well as helping current workers transition to new positions that will utilise their unique human capabilities.

#### A shift in Human Experience

If AI takes over menial tasks and allows humans to significantly reduce the amount of time they need to spend at a job, the extra freedom might seem like a utopia at first glance. However, in order to feel their life has a purpose, humans will need to channel their newfound freedom into new activities that give them the same social and mental benefits that their job used to provide. This might be easier for some people and communities than others. There will likely be economic considerations as well when machines take over responsibilities that humans used to get paid to do. The economic benefits of increased efficiencies are pretty clear on the profit-loss statements of businesses, but the overall benefits to society and the human condition are a bit more opaque.

#### **Global Regulations**

While our world is a much smaller place than ever before because of technology, this also means that AI technology that requires new laws and regulations will need to be determined among various governments to allow safe and effective global interactions. Since we are no longer isolated from one another, the actions and decisions regarding artificial intelligence in one country could adversely affect others very easily. We are seeing this already playing out, where Europe has adopted a robust regulatory approach to ensure consent and transparency, while the US and particularly China allows its companies to apply AI much more liberally.

#### **Accelerated Hacking**

Artificial intelligence increases the speed of what can be accomplished, and, in many cases, it exceeds our ability as humans to follow along. With automation, nefarious acts such as phishing, delivery of viruses to software and taking advantage of AI systems because of the way they see the world, might be difficult for humans to uncover until there is a real quagmire to deal with.

#### **AI Terrorism**

Similarly, there may be new AI-enabled form of terrorism to deal with: From the expansion of autonomous drones and the introduction of robotic swarms to remote attacks or the delivery of disease through nanorobots. Our law enforcement and defence organisations will need to adjust to the potential threat these present.

It will take time and extensive human reasoning to determine the best way to prepare for a future with even more artificial intelligence applications to ensure that even though there is potential for adverse impacts with its further adoption, it is minimised as much as possible.

As is the case with any disruptive event, these aren’t easy situations to solve, but if we still have humans involved in determining solutions, we will be able to take advantage of the many benefits of artificial intelligence while reducing and mitigating the negative impacts.

#### **Social manipulation**

Social media through its autonomous-powered algorithms is very effective at target marketing. They know who we are, what we like and are incredibly good at surmising what we think. Investigations are still underway to determine the fault of Cambridge Analytica and others associated with the firm who used the data from 50 million Facebook users to try to sway the outcome of the 2016 U.S. presidential election and the U.K.'s Brexit referendum, but if the accusations are correct, it illustrates AI's power for social manipulation. By spreading propaganda to individuals identified through algorithms and personal data, AI can target them and spread whatever information they like, in whatever format they will find most convincing—fact or fiction.

#### **Invasion of privacy and social grading**

It is now possible to track and analyse an individual's every move online as well as when they are going about their daily business. Cameras are nearly everywhere, and facial recognition algorithms know who you are. In fact, this is the type of information that is going to power China's[social credit system](https://www.fastcompany.com/40563225/chinas-terrifying-social-credit-surveillance-system-is-expanding) that is expected to give every one of its 1.4 billion citizens a personal score based on how they behave—things such as do they jaywalk, do they smoke in non-smoking areas and how much time they spend playing video games. When Big Brother is watching you and then making decisions based on that intel, it’s not only an invasion of privacy it can quickly turn to social oppression.[[8]](#endnote-8)

#### **Misalignment between our goals and the machine’s**

Part of what humans value in AI-powered machines is their efficiency and effectiveness. But, if we aren’t clear with the goals we set for AI machines, it could be dangerous if a machine isn’t armed with the same goals we have. For example, a command to “Get me to the airport as quickly as possible” might have dire consequences. Without specifying that the rules of the road must be respected because we value human life, a machine could quite effectively accomplish its goal of getting you to the airport as quickly as possible and do literally what you asked but leave behind a trail of accidents.

#### **Discrimination**

Since machines can collect, track and analyze so much about you, it’s very possible for those machines to use that information against you. It’s not hard to imagine an insurance company telling you you’re not insurable based on the number of times you were caught on camera talking on your phone. An employer might withhold a job offer based on your “social credit score.”

Any powerful technology can be misused. Today, artificial intelligence is used for many good causes including to help us make better medical diagnoses, find new ways to cure cancer and make our cars safer. Unfortunately, as our AI capabilities expand, we will also see it being used for dangerous or malicious purposes. Since AI technology is advancing so rapidly, it is vital for us to start to debate the best ways for AI to develop positively while minimizing its destructive potential.[[9]](#endnote-9)

For all processing of personal data using AI systems, controllers11 need to rely on one of the six legal bases for processing set out under Article 6(1) of the GDPR. Most commonly controllers rely on consent, legitimate interests, legal obligation or contractual necessity. 12 An appropriate legal basis should be established during both the training phase and the use phase.[[10]](#endnote-10)

# What should organisations do to ensure that they are being responsible with AI and the wider use of data in general

At the moment, when it comes to AI, companies are expected to self-regulate. This means they must create and implement their own Responsible AI guidelines. Companies like Google, Microsoft and IBM all have their own guidelines. One issue with this is that the principles of Responsible AI may be applied inconsistently across industries. Smaller companies may not even have the resources to create their own.

A potential solution would be for all companies to adopt the same guidelines. For example, the European Commission recently published the Ethics guidelines for trustworthy AI. This details 7 key requirements that AI should have to be considered trustworthy. Using these guidelines would help companies ensure their AI systems meet the same criteria. The real question is — can we trust companies to regulate themselves?

The 2020 State of AI and Machine Learning report[[11]](#endnote-11) included responses from 374 organizations working with data/AI. 75% of the organizations said that AI was a critical part of their business. However, only 25% of them said that fair AI was important. This suggests that the answer is no; no we can not trust them. For common guidelines to be effective, they must also be enforced. In other words, guidelines must become laws/regulations, and companies must face penalties for not abiding by them.

Word Count [ 2285 ]

# References

1. <https://towardsdatascience.com/what-is-responsible-ai-548743369729> [↑](#endnote-ref-1)
2. <https://www.analyticsinsight.net/is-facial-recognition-stooping-down-with-unauthorized-surveillance/> [↑](#endnote-ref-2)
3. <https://www.newsweek.com/stephen-hawking-artificial-intelligence-warning-destroy-civilization-703630> [↑](#endnote-ref-3)
4. <https://www.forbes.com/sites/bernardmarr/2018/11/19/is-artificial-intelligence-dangerous-6-ai-risks-everyone-should-know-about/> [↑](#endnote-ref-4)
5. <https://searchenterpriseai.techtarget.com/definition/machine-learning-bias-algorithm-bias-or-AI-bias> [↑](#endnote-ref-5)
6. <https://www.itpro.co.uk/automation/30463/gartner-by-2020-ai-will-create-more-jobs-than-it-eliminates> [↑](#endnote-ref-6)
7. <https://www.itpro.co.uk/automation/30463/gartner-by-2020-ai-will-create-more-jobs-than-it-eliminates> [↑](#endnote-ref-7)
8. <https://www.fastcompany.com/40563225/chinas-terrifying-social-credit-surveillance-system-is-expanding> [↑](#endnote-ref-8)
9. <https://www.forbes.com/sites/bernardmarr/2018/11/19/is-artificial-intelligence-dangerous-6-ai-risks-everyone-should-know-about/?sh=3d1219d82404> [↑](#endnote-ref-9)
10. <https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf> [↑](#endnote-ref-10)
11. [https://www.stateof.ai/](https://d.docs.live.net/f0c0205b2f9e5eeb/Documents/Data%20Academy%20Reference/Week%207%20-%20TensorFlow/Responsible%20AI.docx) [↑](#endnote-ref-11)