**RESPONSIBLE AI**

Responsible AI is a governance framework that documents how a specific organization is addressing the challenges around artificial intelligence (AI) from both an ethical and legal point of view. Resolving ambiguity for where responsibility lies if something goes wrong is an important driver for responsible AI initiatives.

**AI Fails**

## **Fail: Microsoft’s AI Chatbot Corrupted by Twitter Trolls**

Microsoft made [big headlines](https://www.theverge.com/2016/3/23/11290200/tay-ai-chatbot-released-microsoft) when they announced their new chatbot. Writing with the slang-laden voice of a teenager, Tay could automatically reply to people and engage in “casual and playful conversation” on Twitter.

Tay grew from Microsoft’s efforts to improve their “conversational understanding”. To that end, Tay used machine learning and AI. As more people talked with Tay, Microsoft claimed, the chatbot would learn how to write more naturally and hold better conversations.

Microsoft won’t say exactly how the algorithms worked, of course. Perhaps because of what happened next.

Less than 24 hours after Tay launched, internet Trolls had thoroughly “corrupted” the chatbot’s personality.

By flooding the bot with a deluge of racist, misogynistic, and anti-Semitic tweets, Twitter users turned Tay – a chatbot that the Verge described as “[a robot parrot with an internet connection](https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist)” – into a mouthpiece for a terrifying ideology.

Microsoft claimed that their training process for Tay included “relevant public data” that had been cleaned and filtered. But clearly, they hadn’t [planned for failure](http://lexa.ly/ust3e), at least not this kind of catastrophe.

After a cursory effort to clean up Tay’s timeline, Microsoft pulled the plug on their unfortunate AI chatbot.

## **Fail: IBM’s “Watson for Oncology” Cancelled After $62 million and Unsafe Treatment Recommendations**

No AI project captures the “moon-shot” attitude of big tech companies quite like Watson for Oncology. In 2013, IBM partnered with The University of Texas MD Anderson Cancer Center to develop a new “Oncology Expert Advisor” system. The goal? Nothing less than to cure cancer.

The first line of the [press release](https://www-03.ibm.com/press/us/en/pressrelease/42214.wss) boldly declares, “MD Anderson is using the IBM Watson cognitive computing system for its mission to eradicate cancer.” IBM’s role was to enable clinicians to “uncover valuable insights from the cancer centre's rich patient and research databases.”

In July 2018, Stat News reviewed internal IBM documents and found that IBM’s Watson was making erroneous, downright dangerous cancer treatment advice.

[According to StatNews](https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/), the documents (internal slide decks) largely place the blame on IBM’s engineers. Evidently, they trained the software on a small number of hypothetical cancer patients, rather than real patient data.

The result? Medical specialists and customers identified “multiple examples of unsafe and incorrect treatment recommendations,” including one case where Watson suggested that doctors give a cancer patient with severe bleeding a drug that could worsen the bleeding.

In February 2017, [Forbes reported](https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/) that MD Anderson had “benched” the Watson for Oncology project. [A special report](https://www.utsystem.edu/sites/default/files/documents/UT%20System%20Administration%20Special%20Review%20of%20Procurement%20Procedures%20Related%20to%20UTMDACC%20Oncology%20Expert%20Advisor%20Project/ut-system-administration-special-review-procurement-procedures-related-utmdacc-oncology-expert-advis.pdf) from University of Texas auditors said that MD Anderson had spent more than $62 million without reaching their goals.

## **Fail: Apple’s Face ID Defeated by a 3D Mask**

Apple released the iPhone X (10? Ten? Eks?) to mixed, but [generally positive](https://www.wired.com/2017/11/review-iphone-x/) reviews. The phone’s shiniest new feature was [Face ID](https://www.apple.com/iphone/#face-id), a facial recognition system that replaced the fingerprint reader as your primary passcode.

[Apple said](https://support.apple.com/en-ca/HT208108) that Face ID used the iPhone X’s advanced front-facing camera and machine learning to create a 3-dimensional map of your face. The machine learning/AI component helped the system adapt to cosmetic changes (such as putting on make-up, donning a pair of glasses, or wrapping a scarf around your neck), without compromising on security.

But a week after the iPhone X’s launch, hackers were already claiming to beat Face ID [using 3D printed masks](https://www.wired.com/story/hackers-say-broke-face-id-security/). Vietnam-based security firm [Bkav](http://www.bkav.com/) found that they could successfully unlock a Face ID-equipped iPhone by gluing 2D “eyes” to a 3D mask. The mask, made of stone powder, cost around $200. The eyes were simple, printed infrared images.

Bkav’s claims, outlined in a [blog post](http://www.bkav.com/dt/top-news/-/view_content/content/103968/bkav%EF%BF%BDs-new-mask-beats-face-id-in-twin-way-severity-level-raised-do-not-use-face-id-in-business-transactions), gained widespread attention, not least because Apple had [already written](https://support.apple.com/en-ca/HT208108) that Face ID was designed to protect against “spoofing by masks or other techniques” using “sophisticated anti-spoofing neural networks”.

Not everyone was convinced by Bkav’s work. Publications such as Wired had already [tried and failed](https://www.wired.com/story/tried-to-beat-face-id-and-failed-so-far/) to beat Face ID using masks. And Wired’s [own article](https://www.wired.com/story/hackers-say-broke-face-id-security/) on Bkav’s announcement included some scepticism from Marc Rogers, a researcher for security firm [Cloudflare](https://www.cloudflare.com/en-ca/). But the work – and this glimpse into the weakness of AI – is fascinating.

## **Fail: Amazon Axes their AI for Recruitment Because Their Engineers Trained It to be Misogynistic**

Artificial intelligence and machine learning have a huge bias problem. Or rather, they have a huge problem with bias. And the launch, drama, and subsequent ditching of Amazon’s AI for recruitment is the perfect poster child.

Amazon had big dreams for this project. As one Amazon engineer [told The Guardian in 2018](https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine), “They literally wanted it to be an engine where I’m going to give you 100 résumés, it will spit out the top five, and we’ll hire those.”

But eventually, the Amazon engineers realized that they’d taught their own AI [that male candidates were automatically better](https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine).

How did this AI fail happen? In short, Amazon trained their AI on engineering job applicant résumés. And then they benchmarked that training data set against current engineering employees.

Now, think about who applies for software engineering jobs. And who is most likely to be currently employed in software engineering? That’s right: [white men](https://www.seattletimes.com/business/amazon/amazon-more-diverse-at-its-warehouses-than-among-white-collar-ranks/).

So, from its training data, Amazon’s AI for recruitment “learned” that candidates who seemed whiter and more male were more likely to be good fits for engineering jobs.

That’s the short version – the full story is even more painful. Our article on [bias in AI and machine learning](https://www.lexalytics.com/lexablog/bias-in-ai-machine-learning) has more.

## **Fail: Amazon’s Facial Recognition Software Matches 28 U.S. Congresspeople with Criminal Mugshots**

Amazon’s AI fails don’t stop there. In 2018, the American Civil Liberties Union showed how Amazon’s AI-based Rekognition facial recognition system

[According to the ACLU](https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28), “Nearly 40 percent of Rekognition’s false matches in our test were of people of colour, even though they make up only 20 percent of Congress.”

In fact, that’s not even the first time someone’s proven that Rekognition is racially biased. [In another study](https://www.media.mit.edu/publications/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification/), University of Toronto and MIT researchers found that every facial recognition system they tested performed better on lighter-skinned faces. That includes a [1-in-3 failure rate](https://www.lexalytics.com/lexablog/bias-in-ai-machine-learning) with identifying darker-skinned females. For context, that’s a task where you’d have a 50% chance of success just by guessing randomly.

This is, of course, horrifying. It’s not even an “AI fail” so much as a complete failure of the systems, people and organizations that built these systems.

I wish I could say that, faced with incontrovertible proof that they did a bad thing, Amazon did what they needed to fix their AI bias. But the story doesn’t end here. Law enforcement agencies [are already trying to use tools like Rekognition](https://www.theguardian.com/technology/2017/dec/04/racist-facial-recognition-white-coders-black-people-police) to identify subjects. And despite these demonstrated failures – it’s algorithmic racism, really – [Amazon isn’t backing down on selling Rekognition](https://www.lexalytics.com/lexablog/bias-in-ai-machine-learning).

Seriously, just read this article from The Guardian: [How white engineers built racist code – and why it’s dangerous for black people](https://www.theguardian.com/technology/2017/dec/04/racist-facial-recognition-white-coders-black-people-police)

# **Art. 22 GDPR-Automated individual decision-making, including profiling**

The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

Paragraph 1 shall not apply if the decision:

* + is necessary for entering into, or performance of, a contract between the data subject and a data controller;
  + is authorised by Union or Member State law to which the controller is subject, and which also lays down suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests; or
  + is based on the data subject’s explicit consent.

In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in [Article 9](https://gdpr-info.eu/art-9-gdpr/)(1), unless point (a) or (g) of [Article 9](https://gdpr-info.eu/art-9-gdpr/)(2) applies and suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests are in place.

## **3 steps to ensure AI is served in a responsible way**

So, where do we go from here to ensure that AI is serving our society in a healthy and responsible way? Organisations must think of AI technology in a holistic way – understanding where AI sits in the value chain and creating the right structures to ensure long-term governance by:

1. Establishing internal governance, for example by an objective review panel, that is diverse and that has the knowledge to understand the possible consequences of AI infused systems. A key success factor is leadership support and the power to hold leadership accountable.
2. Ensuring the right technical guardrails, creating quality assurance and governance to create traceability and auditability for AI systems. This is an important part of every organisation’s toolkit to allow operational and responsible AI to scale.
3. Investing more in their own AI education and training so that all stakeholders – both internal and external – are informed of AI capabilities as well as the pitfalls.