The Ethical Al Handbook By Utopia Analytics





Introduction

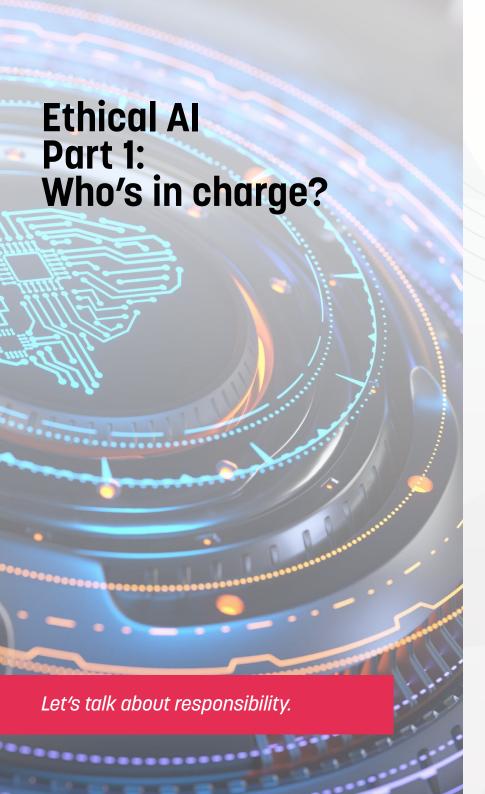
Artificial Intelligence (AI) used to be the stuff of science fiction, but a rapid technological shift has seen AI systems become an essential part of the modern-day world. AI underpins many of the technologies that we interact with on a daily basis and is widely used across a variety of sectors, from e-commerce and finance to healthcare and gaming.

The advancement of AI systems, along with their customizability and accessibility, means they can be deployed in a variety of ways and are perfectly placed to free human staff from the most mundane and repetitive tasks so they can focus on more value-driving work. This can help businesses increase their productivity and profitability while making smarter and more calculated decisions, regardless of the sectors they operate in.

But despite the growing use of AI systems and technology, there's still a general misunderstanding of how they operate.

Research into AI continues to highlight concerns surrounding privacy, surveillance, discrimination and security, and as AI systems more commonplace in our everyday lives, the implications of this technology also needs to be considered alongside the opportunities.

That's why we created the Ethical AI framework, which places principles such as transparency, honesty and non-discrimination at the heart of the way we operate at Utopia Analytics, as well as the services that we build. The Ethical AI framework covers ten key principles, all of which must be considered throughout the development processes of AI systems.



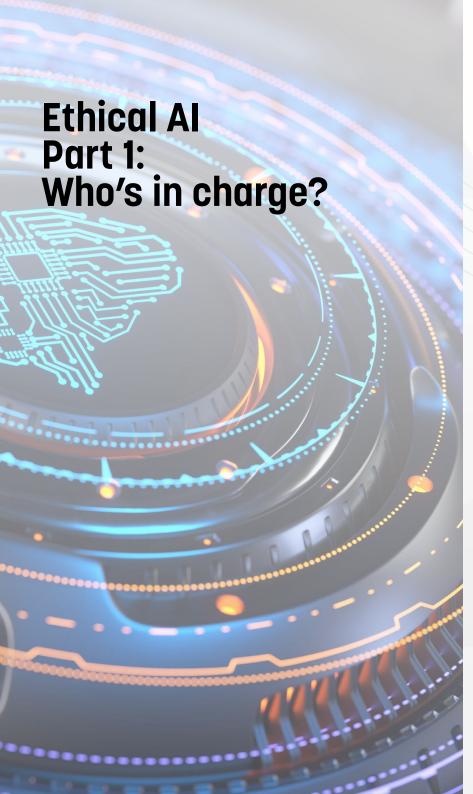
In the airline industry, it's easy to understand how the individual roles and responsibilities of airline staff ensures everyone stays safe. Although Artificial Intelligence (AI) is a relatively new field, the structure of responsible parties and how they work together is quite similar to our airline industry analogy, where the AI could be compared to a plane, where safety and responsibility fall to the pilots and engineers who fly and maintain the aircraft.

But who is ultimately responsible for the decisions taken by AI software? Defining who is in charge of artificial intelligence has prompted a lot of discussion. Experts agree that legislation lays the foundation for the ethical use of AI. In regulatory terms, AI should be seen like any other automated system. And contrary to common fears, there is nothing inherently dangerous or unpredictable about the way AI works.

Right now, machine learning is the most advanced form of Al. Yet despite its powerful abilities, it cannot invent anything creative or new that doesn't already exist beyond the data provided. The safety issues of this kind of Al lie mostly in the collection of personal data and the technologies that have implications for physical interactions with humans, such as machines, cars or aircraft.

One of the most common questions we're asked about machine learning is: 'who's responsible for the resulting actions of an AI?'

Let's return to our airline industry analogy. It's an industry that's well-regulated due to the obvious safety issues surrounding it. Staff onboard the aircraft, as well as its manufacturers, are ultimately responsible for an aeroplane's safe flight from point A to point B. Pilots, engineers, cabin crew and airport staff each have their own areas of expertise and responsibilities, while manufacturers of the aircraft hire professionals that are responsible for safely building the various components of the aeroplane.



Al isn't much different. It's a technology built and used by humans, just like aeroplanes. And similar to aeroplanes, Al applications don't represent a lot of physical risk to humans. This is mostly due to the still-limited application areas of machine learning compared to the technology running on CPUs.

Al could be compared to a plane, where safety and responsibility fall to the pilots and engineers who fly and maintain the aircraft.

The responsibility of AI decision making should work like this: legislators and authorities should define the boundaries for collecting and using data, and what kind of certificates are needed to be eligible to produce AI tools, just like how they would regulate any other tools or products that significantly affect human lives.

For AI, there are three responsible parties:

- 1. The field experts who perform their profession and at the same time provide training data for machine-learning-based Al.
- 2. The AI model and product developers, who need professional skills in understanding how different algorithms behave on different kinds of data, as well as in different types of AI products.
- 3. The users of the AI tool who should follow the user instructions in order to get the expected behaviour of the AI decision-making system.

In order for AI to act ethically, all of these parties must operate transparently and with open communication channels, ensuring that any data fed to AI is trustworthy and of the highest quality.



Al is a powerful tool. For content moderation in particular, it can be used as a force for good by removing hate speech or inflammatory content before it's published, stopping the damage it could cause to readers.

But there are plenty of ways that AI can be misused too. As machine-learning-based AI can mimic the publishing policies of individual communities with high accuracy, it can also be used irresponsibly to silence critical voices if the moderation policy so wishes.

As a text analytics service provider, we ensure that the companies using Utopia AI are in full control of how they utilize it on their platforms. The Utopia AI can adapt to all languages and is able to learn any unique moderation policy in just two weeks. But powerful tools need a responsibly built, well-defined, solid foundation.

While the users of Utopia AI enjoy the freedom of defining their own moderation policies, it's our responsibility to always act ethically within the process of defining this base. Since we understand that definitions of ethics are highly subjective – especially in the field of moderation – we have been working towards achieving a mutual understanding of this not just internally, but with our customers too.

As a text analytics service provider, we ensure that the companies using Utopia AI are in full control of how they utilize it on their platforms.

All things considered, we believe that the United Nation's Universal Declaration of Human Rights (UDHR) is the most suitable for safeguarding the ethical use of Al. In part, because we believe it offers the most comprehensive framework, it's also been



ratified by more countries, compared to other similar declarations, like the European Convention on Human Rights (ECHR).

The Declaration of Human Rights, which was adopted in 1948 by the United Nations General Assembly, consists of 30 articles affirming an individual's rights, as elaborated in subsequent international treaties, economic transfers, regional human rights instruments, national constitutions, and other legal codes and standards.

The Declaration fits well with Utopia's approach to human-first content moderation, which is to focus on the protection of human rights online. By following these guidelines, we're able to safeguard an extensive variety of rights, such as the right to information and the right to participation, without limiting it solely to freedom of speech.

Ethical Al Part 3: Shared responsibility

Human rights breaches as grounds for termination.

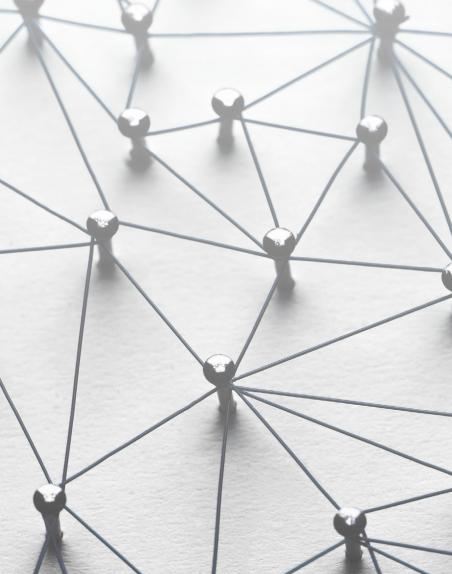
We've already discussed how powerful AI products require a solid foundation that keeps ethics in place for both developers and users. For Utopia, the foundation we use with our customers is the United Nation's Universal Declaration of Human Rights (UDHR). The 30 articles it contains which affirm an individuals' rights act as the backbone for our approach to ethical issues created in the moderation of online content.

Because the ethics of what we do is so important to Utopia, we have somewhat unconventionally included the UDHR in our company service contract as grounds for termination. If a party fails to comply with UDHR, the cooperation will be terminated with immediate effect. This means that anyone using Utopia AI is obliged to comply with the applicable framework set out in the UDHR, while conducting their business in accordance with the highest ethical standards. This ensures that both Utopia and its customers trust each other to act responsibly, and have a clear guide as to what constitutes ethical behaviour.

In the process of keeping the AI model up-to-date, our customers provide us with all of the human moderation decisions that define its moderation policy. Utopia AI learns from these examples while Utopia provides all of its automated moderation decisions for the customer. Both parties are always free to inspect and analyse the quality of each other's actions for transparency.

In the end, the common goal is to protect online discussions without violating freedom of speech and to share responsibility with all parties involved, including the AI service provider, the customers, their users and their content moderators.

Ethical Al Part 3: Shared responsibility



Because the ethics of what we do is so important to Utopia, we have somewhat unconventionally included the UDHR in our company service contract as grounds for termination.

Has our approach worked? Without a doubt.

The Declaration has been in our service contract since the launch of Utopia Al Moderator in May 2016, and has succeeded in helping all involved parties maintain responsibility for the ethical implementation of Utopia Al.

By using something as recognised and detailed as the UDHR, we can start any customer relationship with a clear understanding of what constitutes an ethical foundation of the moderation which will follow. Beyond that, it also removes any unintentional bias that might be introduced if a company was to create its own definition of what ethical behaviours might be.

Ethical AI Part 4: Democracy

It's fundamental to the way we operate that we don't add any layer of interpretation between the online content and the analysis of it, so that there is no opportunity for bias to enter into the workflow.

Imagine if there was an IT company that could access every single piece of usergenerated content on the internet the moment it was submitted. Then, imagine what could happen if that company used a powerful AI to moderate all of that content before it goes live.

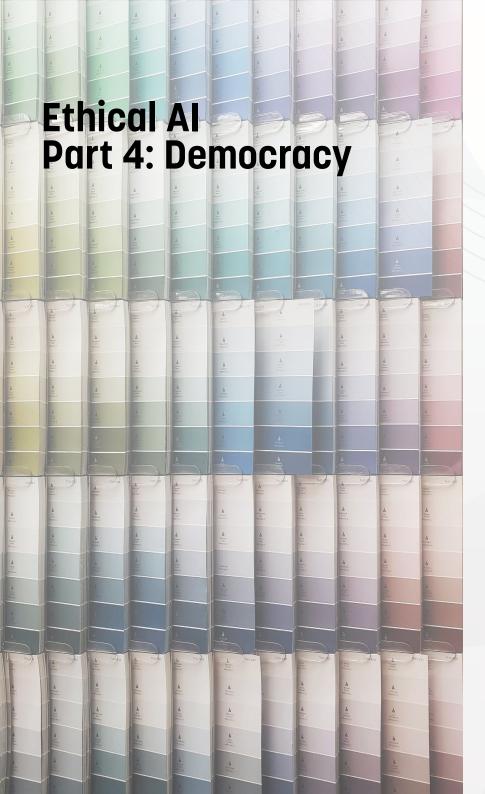
To many people, this is a worrying scenario. And there's a good reason for that.

Modern machine-learning-based AI tools are so powerful that they learn to mimic human behaviour through the process of decision-making. The decisions made by these AI moderation systems heavily rely on training data. As an example, if the training data tells the AI to filter any and all voices that are critical of governments, those voices will be silenced.

So it's important to look at how humans define the moderation policy that's used for their Al.

Utopia Analytics provides real-time moderation services for many of the world's largest online platforms and organisations. The Utopia AI that they use is based on cutting-edge machine learning, while our text analytics products can understand the semantic meaning of any language in the world. It's fundamental to the way we operate that we don't add any layer of interpretation between the online content and the analysis of it, so that there is no opportunity for bias to enter into the workflow.

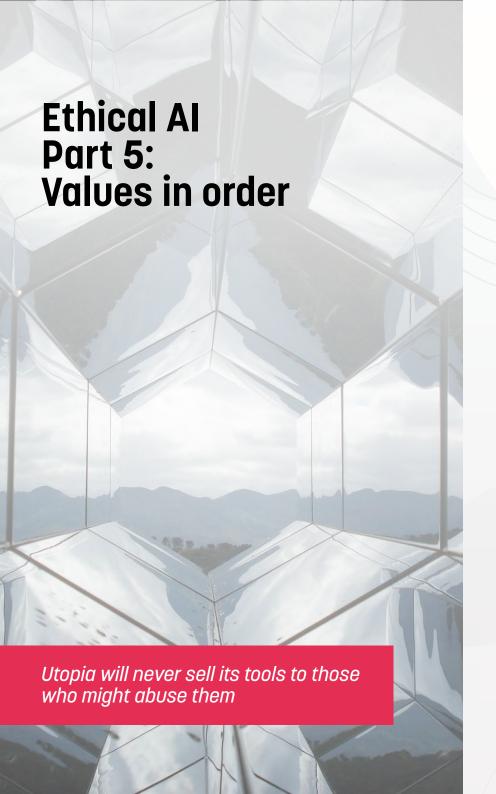
Al moderation and freedom of expression.



With this in mind, it's easy to understand why Utopia has never chosen to define the moderation policy for any of our customers' online services. That is and will always be the customers' responsibility: every company defines itself and how users should behave on their platform, while still complying with international and local legislation.

If there is any doubt that a current Utopia customer or a potential customer would violate human rights through moderation, the moderation service will be immediately terminated or never provided in the first place. Our contracts say that if a party breaches the Universal Declaration of Human Rights, cooperation will be terminated.

We believe it is crucial that people with a diverse range of views and beliefs are able to access the internet, as long as their messaging doesn't hurt anybody. We also believe that no single person or company should be granted the global power to decide what is and what isn't acceptable to say online.



Is it okay to sell the Utopia AI Moderator service to a company that might be in the news for negative reasons, or may not have exactly the same ethical values as we do? This is a question that arises now and then in Utopia's internal discussions and it's an important one.

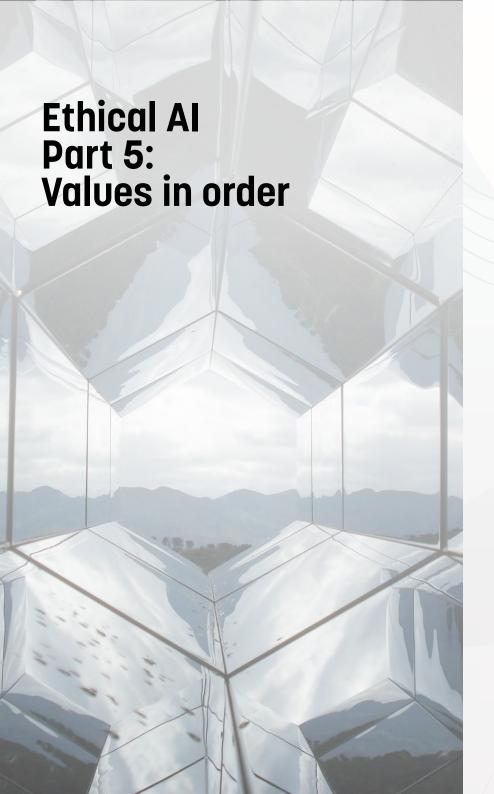
Utopia's Text Analytics Platform is able to understand the semantic meaning in any language in the world, and Utopia AI Moderator learns each customer's moderation policy from content samples moderated by humans. This is called training data, as it helps the AI to understand the different kinds of historic content moderation decisions that have been made, so that the decisions made by the Utopia AI echoes human decision making - but faster and at scale.

It's always the Utopia customer's duty to provide the training data. If human moderation has been biased or has acted in violation of human rights principles, the AI model would learn this behaviour, resulting in biased or dangerous AI behaviour. On top of that, the AI model is more effective than a human moderator can ever be: it can make all the unwanted opinions and voices vanish into cyberspace.

The answer to the initial 'sell/don't sell' question, therefore, depends very much on the particular reason(s) that the company in question wants to use Utopia AI for.

For example, online marketplaces often use Utopia Al Moderator for moderating classified sales ads. Given the nature of marketplaces, it's unlikely that a moderation policy would ever violate anyone's freedom of expression, so it's usually the case that we can offer our services.

However, if Utopia Al Moderator is needed to moderate news comments, the evaluation process needs to be done more carefully. We support freedom of



expression and other human rights, but if there is a risk that a potential customer would, for one reason or another, violate the United Nations' Universal Declaration of Human Rights (the UDHR), we'd never sell to that customer.

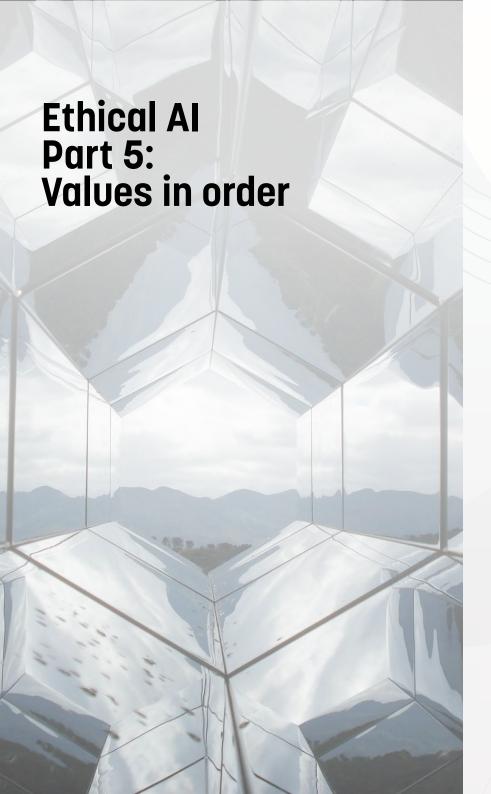
There's a good reason why we take great care when it comes to news comment moderation in particular. In the past, Utopia has been asked to build AI models that could moderate comments in a way that would remove certain types of political opinions. We've always declined such requests, instead offering models that, rather than removing political-related content, would remove content related to hate speech, grooming and other inappropriate behaviour which would not impact the expression of political views.

If a potential customer would [...] violate the United Nations' Universal Declaration of Human Rights (the UDHR), we'd never sell to that customer.

Sometimes it's clear from the start of communications with a potential customer that they are simply not committed to human rights. In cases such as these, the answer to their questions is always an easy one: we're sorry, but we're unable to provide the service you need.

But sometimes the situation is more ambiguous, where a potential customer unknowingly breaches the UDHR. Under those circumstances, Utopia communicates openly with the customer and reminds them that UDHR breach is grounds for termination of the service contract with immediate effect.

Utopia's service contracts usually last for a number of years, so it's important to continuously monitor the human rights situation in the moderation. Our moderation



tools are offered as a service, which allows us to continuously monitor the training data passed from the client for suspicious activity. If things start to go wrong, the termination process is easy for Utopia because we will always follow UDHR, and the stipulation is written into every single one of our contracts.

This is because as a text analytics company, Utopia needs to be just as trustworthy as the most qualified medical practitioners. Quality, safety and ethics always have to come first, which means that we cannot work with companies that don't give these values the same importance.

Our moderation tools are offered as a service, which allows us to continuously monitor the training data passed from the client for suspicious activity.

It may sound unusual for Utopia to place such importance on ethical values in a technology landscape where many platform holders claim no responsibility for the content shared on them. But if we are to be trusted by both our customers and their users to always moderate in a fair and ethical way, we believe it is essential that we model these behaviours in the hope that others in the industry will follow.

Ethical Al Part 6: No prejudice

Our Utopia AI models moderate each message or comment solely by the content and context, no matter who created it. There have been times when we've been asked to build AI moderation models that would prefer certain gender's comments simply because "their comments are better!" Similarly, we've been asked to build AI moderation models that would judge a writer's latest comments based on their previous communications and online behaviour.

Business-wise, we understand the rationale behind both of these requests. If these 'better comments' generate higher levels of engagement and traffic, they're going to generate a larger number of impressions and clicks for any ads on the website, ultimately generating more revenue for the publisher. This is both the strength and the weakness of many social media platforms.

Of course, every company and online service provider has the right and responsibility to decide what kind of comments are accepted on their online service and society.

But if a company wishes to have Utopia AI onboard, the publishing decisions must be made through a process that respects freedom of speech and which isn't prejudiced.

It's unlikely that any company offering moderation tools will set out to use deliberately prejudiced models. However, traditional tools for moderation don't understand the semantic meaning in the text, instead, relying on lists of flagged words or phrases and other data to decide whether a post is acceptable or not. User modelling - i.e. the user's past behaviour - is one way to increase the quality of such moderation tools.

In contrast, Utopia AI is so powerful that it moderates each message or comment solely by the content and context, no matter who created it. We're not willing to build AI models that do user modelling or enable prejudiced or discriminating moderation for social media communication, ensuring users are treated equally and freedom of expression isn't impeded.

Ethical AI Part 7: Honesty



Our professional team focuses on building unbiased AI models.

The human brain is sometimes referred to as the ultimate black box. Science is constantly providing new discoveries but there are many things we're yet to discover! We may tell ourselves that we know our long-time friends or partners well enough to know what they're thinking, but the truth is most of the time we don't have a clue. Even our own ideas, motives, decisions and reactions tend to remain a mystery to us.

Just like the human brain, AI is often called a black box too. And while it's true that many AI products are complicated and some suffer from a lack of transparency, ultimately they are built by humans and with no secret components, so at least the creators of these AI systems know how they work.

Since AI is technology, there is always an explanation behind the behaviour of the AI model, even if it's acting in certain ways that aren't initially understood. With machine learning models, that explanation usually lies in the data describing human behaviour.

Al can actually help to understand humans better, to see inside the ultimate black box of the human brain.

Let's take bias as an example.

Humans are biased in their thoughts and decision-making, and many of these biases we don't even recognise in ourselves. If we look at the hiring process of a new employee as an example, we may lie about or not even realise the specific factors that led to the hiring manager choosing one candidate over another. Similarly, if you're shopping at a supermarket, there will be biases that make you choose one particular brand or product over another.

Someone might let their political opinions affect the moderation decisions they make, whereas another might have created a set of own guidelines for their moderation

Ethical AI Part 7: Honesty



work. This can result in them declining or favouring certain types of content over others. Something as simple as poor sleep or poor mood is another thing that can influence how content moderation is handled. There can be very reasons for these inconsistencies, but ultimately they're the result of natural human behaviour.

In its most elegant forms (such as advanced and very automated machine learning), AI is based on statistical modelling. Such AI can reveal any human bias and inconsistencies in data, and small errors don't interfere with this kind of statistical AI which learns the general rules – the bigger picture – without getting caught up with infrequent errors.

Al can reveal any human bias and inconsistencies in data, and small errors don't interfere with this kind of statistical Al which learns the general rules - the bigger picture - without getting caught up with infrequent errors.

When dealing with freedom of speech and equality and other fundamental human rights, it is essential that the moderation actions are unbiased. The natural tendencies of humans are not always the fairest ones; that's why monitoring processes and tools are needed to remind ourselves to be equal and unbiased in our decisions.

Unfortunately, this means that an AI model can learn to mimic human biases if it isn't built correctly. Creating unbiased AI models is similar to rocket science; data scientists need to deeply understand the behaviour of the AI algorithms they are using, as well as how different types of data can be modelled, in order to get unbiased results.

Ethical AI Part 7: Honesty



As an example, one crucial component is knowing how user information should be used in AI models to not just achieve equal treatment, but to resemble the reality necessary to make useful decisions too. The world around us is not evenly distributed, which means natural data has huge variations. If an AI model is forced to generate evenly distributed results under these circumstances, the model won't reflect society and therefore, won't work.

While there's no denying that some AI systems right now will be making biased decisions, it's important to remember that such decisions are the outcome of human biases.

Thankfully, such decisions can be prevented through effective data governance and by ensuring that AI frameworks are built in a non-discriminatory way by a diverse team.

Ethical Al Part 8: Equality

We provide the same high level of services for all languages and dialects in the world.

Equality. Transparency. Humans first.

These are just a few of the principles commonly listed in fashionable ethical frameworks for artificial intelligence (AI) as released by different organizations.

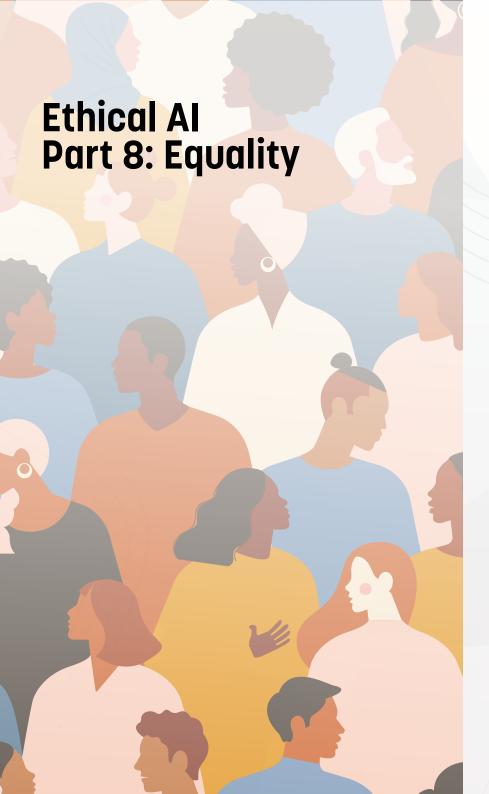
There's nothing wrong with these AI ethics frameworks. They work well for marketing campaigns and for guiding the technical personnel at R&D agencies.

The problem falls specifically within how these principles are implemented in practice: How can we make sure they don't give organizations a false sense that they've made their Al entirely free from potential problems?

Back in 2019, when Utopia finally launched its Ethical AI Manifesto, we had already been implementing all the principles from the launch date of Utopia AI Moderator back in May 2016. The Ethical AI framework was not about creating something new. It was about writing down the Utopia AI way, finding the words for how Utopia's experience of putting ethics into action.

Equality is a noun. One of its definitions is "the condition or state of being the same in number, amount, degree, rank, or quality." For Utopia, this includes providing the same high level of services for all languages and dialects in the world. It's an easy promise to keep, since Utopia AI is language-agnostic.

Unequal access to necessary AI services – for instance, to machine-learning-based text analytics – will generate huge disparities between companies. The innovation



achievement gap in the field of text analytics and other language technologies can also have a corrosive effect on smaller languages, shrinking their influence while the bigger languages keep on growing.

It's an easy promise to keep, since Utopia AI is languageagnostic.

Naturally, equality should also be considered as equal treatment in the analysis of chat messages, news comments, reviews, claims and other sorts of user-generated content. Utopia implements that form of equality by simply not agreeing to build any AI models with prejudice. Utopia AI analyses all items solely according to content and context, no matter who wrote it.



In the physical world, it can be easy to take safety for granted. Most of us recognise the existence of a security structure that ensures everything runs smoothly. If we attend a music festival, there are rules and instructions for everyone to follow, and on top of that, people who make sure the rules aren't being broken without any consequences.

We need to achieve a similar level of safety in the digital world.

We should be able to participate in online gatherings and discussions without the fear of being insulted, humiliated, robbed or attacked.

Responsible digital service providers take the security and wellbeing of their users seriously. They set up terms of use to support the community and its individual members. They watch that the terms are being followed. And if the terms are being violated, they decide on effective consequences that can be implemented without violating anyone's rights.

Nowadays, many digital services have large numbers of users. Whether they're buying or selling, sharing opinions or swapping images, playing or just chatting, traffic in many of these services is so high that looking after the community's wellbeing can take up a lot of time. In fact, sometimes the demands are so high that moderation becomes an impossible task to handle manually. An urgent need exists for advanced digital tools to maintain each service's security, both for the brand and for the users.

If we attend a music festival, there are rules and instructions for everyone to follow, and on top of that, people who make sure the rules aren't being broken without any consequences.



From online marketplaces to news site comment sections to social media services and dating platforms, the service provider must now decide on its own community guidelines for acceptable actions. Utopia AI will learn that policy, and then uses this knowledge to help protect users as well as brands.

Utopia's purpose is to build tools that maintain security and wellbeing in the digital world, and that free humans from mundane tasks, allowing us to focus on the things that really require the attention of our human brains.

Naturally, the policy needs to be uniform and must take account of people's rights. Utopia is not agreeing to build any AI models with prejudice. Utopia AI analyses all items solely according to content and context, no matter who wrote it. Utopia's powerful tools have a solid base.

Ethical Al Part 10: Trust

We do what we promise. And we take responsibility for our algorithms.

All the ethical intentions in the world are worthless without earned trust.

Despite the growing importance and reliance on AI technology, earning the trust of consumers is one of the biggest challenges facing the wider adoption of AI. Factors contributing to the lack of trust in AI systems include concerns surrounding privacy, bias and security to a general misunderstanding of how AI systems function. Levels of trust in AI also <u>vary significantly around the world</u>, suggesting that socio-economic and political factors also need to be taken into consideration when asking consumers to invest their trust in something they can't touch or see.

Trust is an essential part of Al.

Without trust, the decision-making tools powering AI systems cannot deliver the results they were intended to. Honesty and transparency are important contributors to trust in AI and at Utopia Analytics, not only do we consider trust, honesty and reliability as key values of our business, but as cornerstones of our ethical AI services, too. Trust is imbued into every part of our business and the way that we work. After all, if our customers and partners can't trust us to act ethically, how can we ask the same of them?

To ensure neutrality when it comes to decision making, Utopia also never defines a platform's moderation policy; we leave it up to our customers to decide.

<u>The European Commission's Ethics Guidelines for Trustworthy Artificial Intelligence</u> lists seven key requirements that AI systems need in order to be trustworthy, all of which must be continually evaluated and assessed through the AI's lifecycle. These are:

Ethical Al Part 10: Trust

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

We've discussed many of these points, such as accountability, transparency and human agency, in other sections of this handbook. Al models are only as powerful and trustworthy as the training data they are fed, which is why we place such an importance on trust and transparency at an operational level at Utopia. Al must be built and deployed in a responsible manner so we can maximise its benefits to humanity, and this is also the reason why we believe the UDHR is the best foundation to safeguard the ethical use of our Al.

All the ethical intentions in the world are worthless without earned trust.

Whether it's AI moderation or AI-powered claims processing, all of Utopia's AI products have been engineered with the highest standards in mind. Humans always define the task that AI is applied to, ensuring the right ethical considerations are made before the technology is deployed. To ensure neutrality when it comes to decision making, Utopia also never defines a platform's moderation policy; we leave it up to our customers to decide.

As the role that AI systems play in our everyday lives continues to grow, it's essential that both providers and users of these technologies embed trust at the heart of their

Ethical Al Part 10: Trust

operations. Otherwise, issues around consumer attitudes to AI stemming from lack of trust have the potential to be the biggest barrier to consumers experiencing the benefits that AI systems were created to deliver.

The work to build trust into our work begins with the careful screening of prospective customers as part of our sales process, as we don't allow our AI products to be used in situations that may compromise human rights or impede freedom of expression in any way.

More generally, our staff stick to the facts, stick to the timelines, and only make promises that we can keep. This ensures our customers (and their end-users) can always trust us to stick to the same ethical guidelines that we expect them to follow.

We aim to continue this mantra once our work has been delivered and the system has been set up. Our products and AI models are built to last and scale as your business grows. But staff are always on hand for support and maintenance issues with no additional fees.

That's AlaaS (Al as a service). That's responsibility in action. That's the formula of trust, and a key cornerstone of our Ethical Al.

About Utopia Analytics

Founded in 2014, Utopia Analytics comprises text analytics, artificial intelligence and software development experts. With several doctors of science in technology and other key professionals in our team, our heritage is solidly rooted in a background of academic research. Our vision is to change the way user-generated content is utilized and managed.

Utopia Al Moderator and Utopia Al Claim Handler are the company's flagship products. Both can automatically analyse user-generated content in any language using Artificial Intelligence and machine learning.

Utopia's innovative AI-based technologies are already helping create a safe online environment for users of children's Facebook-equivalent Momio, gaming platforms like StarStable and popular online marketplaces such as Swiss Tutti.ch, Finnish Tori.fi as well as Hungarian Jófogás.hu. News comments on Brazilian frontline news channel UOL, Austrian Kronen Zeitung, Dutch Nu.nl and Finnish Helsingin Sanomat are also moderated by Utopia AI.

Find out more at <u>utopiaanalytics.com</u>.