Audio Analytic

Audio for Machine Learning: The Law and your Reputation

Sound recognition requires robust data management.

By Dr Thomas Le Cornu (Data Engineering Manager),
Dr Chris Mitchell (CEO/Founder),
Neil Cooper (VP Marketing)

About us

Audio Analytic is the global leader in intelligent sound recognition. Our cloudless AI technology gives consumer products a sense of hearing beyond just speech and music and, as a result, a greater understanding of context.

Our embedded software platforms (**ai3™** and **ai3-nano™**) are suitable for a wide range of products from smart speakers and video doorbells to smartphones and true wireless earbuds. They have been licensed to some of the world's most prominent consumer technology companies and can be found in many products available today.

Highly-accurate, compact, edge-based AI is only possible because of our world-leading core technology integrated into a specialist ML pipeline.

Machine learning is all about the data, and **Alexandria™** is the world's largest, commercially-exploitable audio dataset for machine learning, with over 30 million labelled recordings across 1,000 sound classes. All of the audio data has been primarily sourced through dedicated data collection campaigns, involving a global network of volunteers or through our purpose-built anechoic sound labs.

We have also designed **AuditoryNET™** — a range of compact, sophisticated, topologically-optimised DNNs and a framework for sound event and acoustic scene recognition, which together model both the acoustic and temporal characteristics of sound. AuditoryNET™ makes use of a dedicated machine learning pipeline which is optimised for sound recognition and includes our patented Loss Function for model training and three levels of data labelling (fine, episodic and weak), and various compression and acceleration techniques.

Contact us

Audio Analytic Ltd
2 Quayside
Cambridge
CB5 8AB
UK

Audio Analytic Ltd
44 Montgomery Street
San Francisco
CA 94104
USA

+44 1223 909 305 | www.audioanalytic.com

Contents

1. Executive Summary

Machine learning (ML) is a data-intensive task, and it is no different when it comes to the specialist branch of sound recognition. Each AI-powered new feature, service or benefit that you bring to your product line introduces significant risks and liabilities if you do not know where the data your organisation uses for training models comes from.

Can you confidently say that you have the explicit permission to use the data you have for the purpose of machine learning? Does that training data infringe upon, or did collecting that training data violate, a consumer's rights to privacy? And do you know that all your data comes with the appropriate licences and doesn't infringe third-party intellectual property rights?

As we highlight in this whitepaper, the impact of being unable to answer these questions potentially includes large fines from regulators, damages claims, claims for an account of profits from intellectual property rights holders, potential exposure to criminal liability in certain jurisdictions, poor long-term stock market performance, and a reduction in customers, among other liabilities.

The challenge is that laws relating to privacy and intellectual property rights around audio data for machine learning are complicated. Plus they are changing quickly as the rest of the world catches up to the role data plays in ML, and the need to protect it. While regulators have been a bit behind the AI/ML curve, they are now starting to ask the above questions around the vast swathes of data that fuel the algorithms.

Companies are expected to secure unambiguous, explicit rights to use data, they must not infringe intellectual property rights, they should ensure traceability, and finally must demonstrate that they can sensitively manage potentially personally identifiable data with the care and attention that it deserves. At the same time, companies must navigate robust, punitive privacy regimes that can vary significantly from country to country; a challenge for any multinational.

As the global leader in sound recognition, we recognise the responsibility that we have towards data, and we take it very seriously. We follow the principles set out in this whitepaper, which means that when we license our technology to the world's leading consumer tech companies, they can confidently empower new products without having to take on potential unknown liabilities.

“The bigger the data set, the better the algorithm, and the better the product for consumers, end of story... right? Not so fast. Be careful about how you get that data set... As the Facebook case shows, how you get the data may matter a great deal.”

Andrew Smith, Director, USA Federal Trade Commission, Bureau of Consumer Protection - April 8th, 2020 (source)

2. Introduction

In the products we buy and the services we use, we can find examples of AI (Artificial Intelligence) improving experiences, processes, or outcomes. Companies, whether they are selling consumer products or enterprise services, like to emphasise their AI credentials because there is a strong demand[1] for products and services that have greater intelligence, personalisation or autonomy, even if the target audience doesn't really understand what that means.

This strong appetite for AI is driving many companies to look at how they can embed more AI into their products. It has been, and continues to be, an exciting time to be working in AI—which is going through a sustained gold rush. Yet, in a hurry to find these killer AI applications, many companies are failing to understand the significant risks and liabilities involved in designing, developing, and training these systems.

The training of AI systems is what we call machine learning (ML). In simple terms, these systems require data as well as the associated algorithms, tools, techniques and expertise to learn patterns. Typically, the more diverse and representative data you have, the better your system performs—and the better the outcome and experience for end-users. Clearly, the companies that can offer the best performance and experience can demand a premium—whether in financial or market positioning terms.

This creates an overwhelming demand for data, and here lie some of the most significant potential liabilities.

Regulators around the world are waking up to the risks of allowing the data demands triggered by the AI gold rush to continue unchallenged. In addition, consumers are becoming savvier as they learn more about what AI is and what it means for them. In turn, regulators recognise that their role is to protect consumers and encourage competition. As the regulators' understanding of the ML process catches up, they are starting to ask probing questions regarding the legal and ethical aspects around data in particular. For example:

- Where did the data come from?
- Do you have permission to use it for the intended purpose?
- Does it infringe upon, or did its collection violate, an individual's rights to privacy?

More generally, increased regulator attention around the world on privacy law is having a significant impact on companies, resulting in fines, restrictions, and damaging brand reputation. Here are a few examples:

- The US Federal Trade Commission (FTC) imposes $5bn penalty and sweeping new privacy restrictions on Facebook.
- France's National Data Protection Commission (CNIL) imposes a financial penalty of €50m against Google for failing to have a lawful basis to process user data.
- South Korea's Personal Information Protection Commission (PIPC) fines Facebook ₩6.7bn ($6.1m) for sharing data without consent.
- The US FTC reached a settlement with Californian-based Everalbum, which deceived consumers about its use of facial recognition technology. This settlement requires the company to delete models & algorithms it developed by using the data obtained unlawfully.
- South Korean regulator fines TikTok ₩186m ($155,000) over mishandling child data

1 - Our recent, independent consumer surveys on the smart home and hearables show a strong positive response to the role of AI.

- Canada's privacy authorities find that Clearview's AI practices are unlawful and recommend the company delete all collected images and biometric facial arrays of individuals in Canada

It is not just the regulators that are putting companies under greater scrutiny. Consumers are also becoming more sensitive to how their data is being used and are taking privacy and respect for their sensitive data very seriously (see section 5 in this whitepaper). As a result, we have seen market-leading companies such as Apple, position consumer privacy and data responsibility as a key defining corporate value.

One of the clearest examples we have seen in recent years of strong, negative consumer responses is when IBM trained a face recognition system using images from Flickr. The company took images uploaded to Flickr under Creative Commons licence conditions and used them to train their system. This mistake resulted in an avalanche of negative media coverage:

- Facial recognition's 'dirty little secret': Millions of online photos scraped without consent (NBC News, March 12th 2019)
- IBM didn't inform people when it used their Flickr photos for facial recognition training (The Verge, March 13th 2019)
- IBM used Flickr photos for facial-recognition project (BBC News, March 13th 2019)
- IBM stirs controversy by using Flickr photos for AI facial recognition (CNET, March 13th 2019).

A year later, IBM's CEO, Arvind Krishna, wrote to Congress to confirm that it had cancelled its facial recognition system. However, the issue has not disappeared. At the time of writing, a class-action lawsuit (Vance v. International Business Machines) is currently working its way through the court system in Illinois.

The issue, we believe, is that decisions over datasets are taken by research teams buried deep within corporate organisations. Their efforts are on proving technical feasibility, and the legal risks may not be known or tracked as this new capability emerges from R&D into product management and marketing. At this point, organisations focus on the performance of features powered by AI systems rather than questioning the data sources. This could be as a result of employing ML practitioners who often come from academic backgrounds—where datasets are readily available—not realising that this is no longer the situation.

Whatever the reason, decisions that are being made deep within companies by researchers and engineers are having serious implications for the companies' finances and brand reputation. As a result, both the technologists leading new ML initiatives and the people ultimately responsible for the company's activity (the board) need to understand and question where their data comes from.

This issue is compounded by the fact that these are complex and relatively new areas of privacy and copyright law. There is often little case law to rely on for highly technical interpretations of how privacy and copyright should be applied to machine learning. The EU recognises these issues and has embedded the principles of privacy by design and privacy by default as the heart of Europe's GDPR.

In this whitepaper, we will look at the legal challenges associated with using audio data to train a sound recognition system, how we expect those challenges to evolve, and what companies need to do to ensure that they are compliant.

3. Audio data rights management for sound recognition is complicated

Audio data and the rights to use it for machine learning purposes are complicated. And the complications and restrictions increase with scale—both from the perspective of technological capability and geographical coverage. For the purpose of this whitepaper, we shall focus on the two broad categories that carry the largest risk of potential liabilities:

- Category 1: Audio data that contains personally identifiable information (PII).
    - For example, an audio recording of somebody saying their name.

- Category 2: Audio data (or other data sources that contain an audio component, such as video) that is covered by copyright or is available under certain licence conditions.
    - For example, video content uploaded to YouTube is made available to other users under certain licence conditions that limit their use of the content to YouTube's platform, and the copyright usually sits with the person who created the content.

A key part of training sound recognition systems is to make sure that the models are trained to recognise certain sounds and ignore others. This means that, during the training process, an ML engineer will need both 'target' and 'non-target' audio data. Both are equally vital to the training process, so all data requires the same level of scrutiny.

There are additional layers of complexity within category two that further increase risk:

- Capturing the real sounds in each environment is critical, and it is, therefore, important to understand the different laws in each location and how it applies to your data.

- If you are recording data in public spaces, it may still be the case that you need a licence, especially if that recording is to be used for 'commercial purposes'.

- If you do have a licence, does the entity that granted it have the rights to do so? For example, in the previous section, we mention the ongoing class-action lawsuit in the US (Vance v. International Business Machines) where IBM's lawyers felt that Creative Commons via Flickr gave them the right to use these images for facial recognition. The lawsuit was filed against them for releasing the Diversity in Faces dataset, which was used by IBM as well as Amazon, Microsoft and Google's parent company Alphabet to improve their facial recognition software. Another interesting case regarding copyright, which included an intermediary was Davidson v. United States. A sculptor successfully sued the US Postal Service for $3.5m for using a picture of his sculpture, which they licensed through an intermediary (in this case Getty Images), without his permission.

- It cannot be assumed that the licence terms are granted in perpetuity; in most cases, they are not. Parties that commercialise data and intellectual property rights will control access through contracts. Likewise, rights secured in relation to personally identifiable data, even where based on consent are neither irrevocable nor perpetual. Rights granted in relation to data do expire with the passing of time and the changing of circumstances. Businesses have a responsibility to manage rights now and in the future. This is relevant whether those rights are granted by or through an organisation or directly by the individual. A central objective with data privacy legislation is to give individuals control over how their personal data is used—this includes giving them the ability to revoke access.

Audio data rights management for sound recognition is complicated

The rights-management side of machine learning is a significant undertaking, especially as datasets need to grow as the technology develops and scales. Each data point—recordings in the case of sound recognition—needs to have a fully traceable audit trail, along with appropriate licences and evidence of the right to use it for the correct purposes. As we shall explain in further sections within this document, future regulations may require companies to produce full audit trails on datasets. This is straightforward if the data pipeline and platforms have been built correctly from the outset, but a significant undertaking if that audit trail doesn't exist.

In the following section, we look at the current regulations around each of these areas in greater depth.

4. Current rights and issues for audio data and sound recognition

4.1 Personal data, rights and audio around the world

The regulations around personal data—and in particular personal audio data—differ around the world, so for simplicity and brevity, we will focus on three example areas: Europe, the US (specifically California), and South Korea. As mentioned in the previous section, personal data regulations apply to all sound recordings and metadata used in training of the sound recognition system.

4.1.1 Europe

Within Europe, privacy is considered a human right. The processing of personal data is governed by GDPR (General Data Protection Regulation, 2016/679), which regulates data processing across the whole of the European Union (EU) and European Economic Area (EEA). Each member state of the EU and EEA has its own data protection supervisory authority, and in some cases, such as Germany, there are regulators for each region within a country overseen by a central regulator. Although the UK has left the EU, GDPR has been retained in UK law. The UK's regulations may diverge from the EU's in the future, but the UK has stated that it is committed to high data protection standards and so any watering down of the obligations implemented in the UK by the GDPR are unlikely.

The primary aim of GDPR is to give individuals control over their personal data and to simplify the regulatory environment for international business. The GDPR established a global standard for data protection. It has a broad reach and applies to data processing by EU established businesses wherever the data is located in the world, and further extends its scope to non-EU companies that employ EU citizens or offer goods or services to individuals or monitor their behaviour.

Processing activities must be justified under a lawful basis, only one of which is consent. If consent is the basis chosen to justify processing, organisations must obtain explicit permission to process the personal data for the specified activity, using language that clearly describes how the data will be used. Such consent must be use-specific, meaning that data collected for one reason can't be used for another purpose and that organisations cannot collect more data than is necessary for the stated purpose. In addition, organisations must make it as easy for individuals to withdraw their consent at any time as it was for them to provide it in the first instance.

GDPR Article 4.1 defines personal data as being "All data related to a person and that allows them to be identified directly or indirectly (...) to one or several specific properties unique to their physical, physiological identity." There is nothing inherent in audio data that automatically exempts it from not being personal information. In fact, France's data protection regulator, the CNIL lists sound recordings of voices as examples of personal data.

Organisations that fail to comply with the GDPR can be fined up to 4% of global annual turnover or €20 million (US $24 million), whichever is greater.

4.1.2 USA

In the US, the California Consumer Privacy Act (CCPA) came into effect on January 1st, 2020. The CCPA changes the way Californian residents can handle their own data as it empowers them with new rights to request businesses to disclose or delete the data they have already collected, or to opt-out completely of third-party data sales.

In late 2020, Californian residents voted for 'Proposition 24' (the Californian Privacy Rights Act (CPRA))—which gives additional protections for sensitive personal information, empowers Californians with further privacy rights and introduces new regulations on businesses that process the personal information of California residents. This new state law, which amends and strengthens the original CCPA legislation, will also see the formation of a California Privacy Protection Agency. CPRA comes into force on January 1st 2023 and will cover data collected from January 2022.

CCPA is often compared with GDPR, but there are subtle, though important differences that need to be understood. While GDPR is focused on creating a 'privacy by default' legal framework for the entire EU, CCPA is about creating transparency in the state's huge data economy and granting clearer rights to its residents.

In other words, while GDPR creates a door for the EU user to lock prior to any data processing, the CCPA creates a window for Californian consumers to open, in order to find out what of their data has already been obtained by a business or sold to a third party.

The CCPA provides California residents with the right to:

- Know what personal data is being collected about them.
- Know whether their personal data is sold or disclosed, and to whom.
- Say no to the sale of personal data.
- Access their personal data.
- Request a business to delete any personal information about a consumer collected from that consumer.
- Not be discriminated against for exercising their privacy rights.

A comment submitted to the California Attorney General's final California Consumer Privacy Act regulations asked if audio recordings are personal information under CCPA and should they be included in the specific pieces produced as part of an access request? The California Attorney General replied "yes" and "yes".

4.1.3 South Korea

The South Korean approach is seen as one of the strictest data protection regimes in the world. The Personal Information Protection Commission (PIPC) is an independent body established under the Personal Information Protection Act (PIPA). The PIPC became the main privacy regulator and enforcer with the amendment of the PIPA in August 2020.

The general approach to privacy regulation in South Korea is to require prior express opt-in consent from the 'data subjects' across most stages of personal information processing (collection, use, and controller-to-controller transfer). For example, unlike in some other countries, a data controller must obtain consent from the data subject in order to collect and use their personal information through a dedicated consent form. Simply having a privacy policy or having the data subject consent to the privacy policy is insufficient.

The South Korean regulators and courts tend to take a rather conservative position in interpreting various legal requirements and definitions under the PIPA. Plus, the regulators are aggressive in enforcing the law, including imposing strict substantive and formality requirements on privacy-related documents, such as consent forms and privacy policies.

Within the South Korean legislation, personal data is defined as:

- Any information that identifies an individual by his or her full name, resident registration

number, image, etc.
- Any information which, by itself, does not identify an individual, but may be easily combined with other information to identify an individual. The ease of combination is determined by reasonably considering the time, cost, technology etc. used to identify the individual and the likelihood the other information can be procured.

Sound is covered under the PIPA as personal information if it either identifies a specific individual by itself or can be easily combined with other information to identify a specific individual.

The reach of PIPA is not limited to South Korea. Although the law does not have a specific provision relating to its extraterritorial jurisdiction, the South Korean regulators routinely reach out to foreign entities which are deemed to be doing business in South Korea, and if a violation is found, have investigated and sanctioned foreign companies (such as Facebook and TikTok). The stated key rights that individuals have in relation to the processing of their personal data are:

- Right of access to data/copies of data.
- Right to rectification of errors.
- Right to deletion.
- Right to object to processing.
- Right to restrict processing.
- Right to data portability.
- Right to withdraw consent.

Given the strict legal requirements and strict interpretive positions taken by the regulators, it is important for a foreign company doing business in Korea to fully comply with the South Korean privacy law requirements, including obtaining due consent from the data subjects before their personal information is collected and used.

4.2 Licenses and audio

Copyright vests automatically in the creator of an original work of authorship that has at least some minimal degree of creativity—which could include audio recordings[2]. Generally speaking, the owner of the copyright has the exclusive right to copy, prepare derivatives works of, distribute, publicly perform, and publicly display the work (and to allow others to exercise those rights). For the purposes of machine learning, it is important that we understand the impact of copyright and licensing. For example, YouTube typically restricts each user's right to view or listen to content on its platform to that user's personal, non-commercial use. Copyright holders, who upload content to YouTube, typically grant other users a licence to access the content that is hosted on the platform only as enabled by YouTube. There are some exceptions to the standard licensing terms, such as Creative Commons, but these do not necessarily grant the required rights for machine learning (see our previous reference to Vance v. International Business Machines).

Although they differ around the world, copyright laws allow for 'fair use' to balance the rights of copyright owners with a range of other rights, interests, and freedoms. However, the exact wording of the legislation differs, as does the judicial interpretation and application. Crucially, 'fair use' does not permit commercial exploitation. Therefore, within an academic setting, the use of copyrighted audio data may be acceptable under non-commercial fair use terms, but if the resultant model were made commercially available, then the data used may no longer be 'fair' to

2 - In some jurisdiction such as South Korea, it is 'neighbouring rights' not 'copyright' that is granted to the creator of a recording. However, under the neighbouring rights, the framework of copyright system or the need for securing a licence would remain the same with copyright.

use. This is a potential area where organisations will be caught out if their focus is on the model and it's capabilities, rather than on the underlying data used to train it.

Within the EU, there are other relevant intellectual property rights to consider. An organisation may have database copyright or a standalone database right. These rights come into existence automatically without any registration requirements. For example, if an organisation has collated sounds in a systematic or methodical way, where the resulting collection constitutes the author's intellectual creation, or has made a substantial investment in obtaining, verifying or presents the contents of a database. As a result, it then becomes possible for infringement of database rights to occur even if other forms of copyright are not infringed. This is a particular risk in the context of AI and ML, where organisations want as much data as possible. Stumbling across a database of sounds and using those sounds could easily constitute extraction or reutilisation of a substantial part of a database, which if protected, will mean an infringing act has been committed.

Rightsholders can give permission to exercise a right—where the holder allows a third party to exercise that right in return for some monetary or non-monetary consideration.

There are many aspects to a licence to use audio data, for example, its defined purpose and whether its royalty-bearing or royalty-free, exclusive or non-exclusive, limited or unlimited use, and commercial or non-commercial. If your machine learning system requires audio recordings that have been created or recorded by other people, then you need to license this data and make sure they have the right to license it to you. If you license audio data from a third party, it carries significant technical risk—which we discuss in a previous whitepaper titled Why Real Sounds Matter for Machine Learning. Additionally, existing licences are not designed for the purpose of machine learning and commercial deployment. Explaining what machine learning entails is not easy to do with people who are unfamiliar with the concept. As a result, significant ambiguity remains over usage rights, and this leads to significant potential liabilities.

To further complicate matters, it is the responsibility of the licensee to have fully understood the licence chain. If you are licensing recordings directly from the copyright holder, then there is one licence to agree. However, if you are licensing audio recordings via a central intermediary, then there are two (or more) licence agreements that are needed to allow you to train an ML model for commercial deployment. Misplaced assumptions that liability ends with the intermediary can leave companies exposed to claims for unauthorised use.

If you are able to agree to licence terms with each rights holder, then the next matter at hand is to determine whether you have a perpetual right or one that carries a limited term-of-use. Any machine learning system built upon a limited-term licence would need to be retrained without that data at some future point in time, which may compromise the reliability of the system. This places a burden on the data platform to provide traceability so that expiring licensing terms can be easily assessed and mitigated. However, it is better to avoid expiring licensing terms in the first place, which is why emphasis should be placed on primary data collection.

A final word of caution related to criminal liability: in certain jurisdictions[3], laws make express reference to how, in certain circumstances, acts of infringement can constitute a criminal offence. We have seen in this section that the rights around audio data are complicated. In the next section, we look at how evolving AI legislation could make these rights yet more complicated.

3 - See as an example, the UK Copyright, Design and Patents Act 1988

5. Potential future rights and issues for audio data as AI legislation evolves

In October 2018, Apple CEO Tim Cook told a data protection conference in Brussels that "We should celebrate the transformative work of the European institutions ... It is time for the rest of the world, including my home country, to follow your lead."

Let us assume, for the purpose of this whitepaper, that the principles and level of regulation proposed by GDPR are adopted by the major economies of the world. How do we see these regulations developing over time?

In February 2020, the European Commission published a whitepaper entitled Artificial Intelligence: A European approach to excellence and trust. In the whitepaper, the authors set out requirements for high-risk AI applications which cover training data, record-keeping, transparency, robustness, accuracy, and human oversight.

Understandably, the European Commission whitepaper focuses on what it deems to be 'high-risk' applications as a priority, as these are the areas where significant damage can occur to citizens of EU countries. However, we believe that traceability ('record-keeping' in the EU proposals), which is a key feature of the GDPR, will permeate into the wider AI industry.

As mentioned in the previous sections of this whitepaper, given the potential for personally identifiable information or copyrighted content to find its way into commercially-available systems, companies will need to protect their investments and adopt detailed record-keeping practices to reduce risk and exposure to potential liabilities.

This approach protects organisations from:

- Regulators challenging them on whether their systems correctly manage personally identifiable information and that models are free from unintended bias.
- Legal challenges over the use of data protected by copyright or other intellectual property rights or breaches of licence conditions.
- Data providers requesting detailed information on how their data is being used.
- Future regulatory changes that widen the definition of 'high-risk'.

Organisations adopting such detailed record-keeping would be able to easily challenge claims, data removal requests, as well as support positive company positioning statements around 'transparency' to a global market where data protection and responsible data management are becoming increasingly important to consumers.

Regulators care about transparency (a core principle under the GDPR) and traceability as it protects civil liberties and fosters public trust and confidence, which in turn benefits companies as it promotes the adoption of new AI-based products and services, driving revenue and healthy competition.

As set out in the introduction, a detailed record-keeping process enables organisations to easily and categorically answer the following three critical questions about data:

- Where did the data come from?
- Do we have permission to use it?
- Does it infringe upon, or did its collection violate, an individual's rights to privacy?

This traceability initiative means that not only will companies need to adopt these approaches within their organisations, but they will also need to enforce adoption throughout their supply chains. In July 2020, the Italian Data Protection Authority (GPDP) fined mobile telecoms company Wind Tre over €16m over GDPR violations. As part of its ruling, the regulator found that Wind Tre did not perform sufficient due diligence on its partners (one of its partners was also fined). As a result, it is clear that regulators expect companies to have conducted sufficient due diligence upon those partners who could have access to large amounts of data.

This further strengthens the case for the universal adoption of transparent record-keeping so that those who can offer full traceability can eliminate risks for large companies facing very large fines in the future.

6. The potential impact of all of this on reputation

For a clear example of how poor reputation data management can negatively affect a business's reputation and result in dramatic consequences, we refer you to the recent WhatsApp privacy policy update. According to reports in the media, users have been flocking to rival platforms such as Telegram and Signal following concerns that the policy change will see sensitive personal information shared with WhatsApp's parent company Facebook.

According to a Pew Research Center survey in June 2019, US consumers do not trust organisations when it comes to their privacy. Sixty-two per cent said that "it is not possible to go through daily life without companies collecting data about them." Further findings among American adults include:

- 81% felt that they had little or no control over the data that companies collect.
- 81% felt that the potential risks of companies collecting data about them outweighed the benefits.
- 79% felt that they were very or somewhat concerned about how companies use the data they collect.
- 75% said that there should be more government regulation of consumer data.

This lack of trust is manifesting itself in purchasing behaviour. A 2020 survey, also conducted by Pew Research Center, found that 52% of American adults (approximately 109m people) had decided not to use products or services due to privacy concerns.

Many media outlets, including established broadcasters and newspapers, digital, professional, semi-professional and even blogs, are focusing on the issue too. As the WhatsApp/Facebook example highlights, companies can ill-afford negative media coverage.

The impact of poor practices is not just the odd negative headline. Data breaches—where private consumer data has found its way into the public domain through mistakes or poor security—are a useful reference point. As reported in Forbes in November 2019, a report from technology research firm Comparitech assessed the impact that data breaches had on share price and found that:

- Share prices of breached companies fall 3.5% on average.
- Six months after a breach, the companies that Comparitech analysed performed worse than they did in the six months prior.
- In the long term, breached companies underperformed the NASDAQ by 8.6% after one year, 11.3% after two years and 15.6% after three years.

So, while it may be tempting to disregard poor data management as just a bit of negative publicity, it is clear from the above research that a poor reputation can have a significant impact on a company—and this can manifest itself financially.

7. The key pillars of a robust data management approach

The companies that can demonstrate that they are acting responsibly, legally, and ethically, are the ones that stand the best chance of building and maintaining consumer trust, as well as avoiding regulator fines and dismissing legal challenges. This calls for a robust approach to managing data for machine learning.

It is important that any approach to data management considers the global reach of products and the shifting trends in data regulation as it applies to data privacy and AI. As such, we have identified the following key pillars of a robust data management approach:

1. **Permission to use**

Training data should only be obtained if you have clear permission to use it for its intended purpose. Copyright holders or contributors must clearly understand what their data is being used for and that permission needs to be expressed clearly through agreed licensing terms.

2. **PII identification and management**

Even though you have permission to use personally identifiable information for the purposes set out in any licence agreement, due care and attention must be placed on recognising, identifying, and managing relevant PII. Any PII that is irrelevant to the machine learning task should be thoroughly removed.

3. **Traceability**

The key to traceability is a robust audit trail so that audio and associated metadata can be traced back to its source, licences and permission, as well as any models that it used to train. Traceability should ensure that it is possible to not just locate the source material but any material that is derived from it. Examples of this kind of modification include augmentation and auralisation, which are complex audio related techniques used within the field of sound recognition.

4. **Limited access privileges**

PII should only be accessed by the people who need it. All data management systems should be secure and locked by default, with ML engineers and researchers only able to access what they need. Certain PII should be stored securely and solely used for traceability and data management. Anybody who has access to any data must be covered by contractual obligations that place data protection at the heart of what they do.

5. **Responsibility**

A dedicated individual or team should be assigned the ability and responsibility to manage data and should be up to date with changes in legislation.

8. Conclusion

Companies need to take responsibility for the data that they collect and use. This is because regulators, content creators and consumers are starting to realise that these wonderful new AI-powered features and capabilities that we find in our homes, phones and earbuds require an enormous amount of data. As a result, they are starting to ask questions that organisations are ill-prepared to answer.

The impact of not understanding whose data they have and whether they have clear permission to use it opens up significant liabilities related to privacy and intellectual property laws. As we have shown, this can lead to fines and long-lasting damage to brand reputation, that in turn has a direct impact on revenue and share price.

However, if you have adopted a robust approach to data management from the outset, as set out in section six, then it is possible to deliver reliable and accurate sound recognition technology that empowers exciting new values, features and applications without infringing privacy and intellectual property rights.

About Audio Analytic

Audio Analytic is the pioneer of AI sound recognition technology. The company is on a mission to map the world of sounds and give machines a compact sense of hearing. By transferring our sense of hearing to consumer products and digital personal assistants we give them the ability to react to the world around us, helping satisfy our entertainment, safety, security, health, wellbeing, convenience, and communication needs.

Learn more: audioanalytic.com