StageZero
TECHNOLOGIES

WHITE PAPER
2023

StageZero's checklist to ensuring privacy compliance globally

stagezero.ai

How to ensure data compliance in AI development | The StageZero Checklist

A critical aspect to consider when developing AI and NLP models is compliance with data regulations. A reliable voice assistant or chatbot requires lots of quality training data, and companies often turn to external data vendors for help. If you are also looking for one, make sure that your chosen data partner utilizes the right tools and methodology to work with sensitive data.

At StageZero, we take regulatory data compliance seriously. We aim to provide our clients with machine learning data that follows the highest security standards and adheres to all necessary international data regulations.

Each time we work with a new data case or segment, we follow these 8 steps to ensure compliance and minimize data vulnerability risks:

1. Initial case analysis and data sensitivity assessment

We analyze each new use case and note down potential solutions and respective considerations. This includes finding answers to questions such as: How will the data be collected? What will the data processing look like? Where will the data be stored? Does it contain personally identifiable information (PII) or biometric data?

PII is data that can be used to identify a person. It covers a broad spectrum of personal details, including name, birth date, username, passwords, credit card information, or social security number. Different from biometric data, PII is changeable.

Biometric data is any data related to human features or characteristics: fingerprints, irises, voice, DNA, or behavioral patterns. One of the most common cases of biometric data usage is iPhone's fingerprint and facial recognition technology. Biometric data is typically highly sensitive information and can present more challenges in complying with privacy laws and regulations.

2. Outline the case and evaluate required regulatory compliance

After initial analysis, the next step is to create an outline for the case with proposed solutions. For example, take a particular customer service use case. We may suggest collecting data from fifty thousand people, including details such as gender and age group, but excluding any unnecessary PII. We use other users to review and validate the data to ensure quality. If the data contains PII, a consent form is needed from the people collecting the data.

Once the initial outline is drawn and solutions suggested, our legal counsel validates corresponding regulatory data use compliance. If needed, solutions are updated.

Data use compliance concerns laws and standards that regulate how organizations collect, store, and manage data. An organization is data compliant if it handles data following the required regulations.

3. Present potential solutions and assess risk

Having evaluated regulatory compliance, we introduce potential solutions to the client. The advantages and disadvantages of each solution in terms of risk are also presented. Depending on the client's legal strategy, they may be more or less risk averse.

Often, it is impossible to avoid risk entirely, especially when data includes PII or biometric characteristics. In those cases, it is up to the customer to decide what risk tolerance level they are comfortable with.

4. Create a data protection impact assessment (DPIA)

We create a DPIA for each new type of case (for example, speech data collection) that can contain PII per our assessment. DPIA is the primary tool to keep the data secure and helps avoid fines should your data be leaked or hacked. DPIA is used to identify and protect against any data privacy vulnerabilities that certain scenarios or activities might cause.

Not all data-related risks can be foreseen or eliminated, but a DPIA gives you an excellent basis to prepare for data protection challenges, set out plans for solutions to address those risks, and evaluate project viability from the get-go. Having a DPIA in place also helps communicate better with your stakeholders regarding data security risks.

If you as a company can show that you followed the best data protection practices and have tried to mitigate risks, the chance of running into legal difficulties or getting a fine is significantly reduced.

Find more info about DPIA and download a template.

5. Define and maintain a data security policy

An additional tool to reduce risk is to maintain a data security policy. This policy indicates how sensitive data should be handled. In other words, it means documenting how data is processed and transferred between collaborating parties and internally.

Each company should have its own data security policy. We also recommend having a general policy for handling data and a section connected to the AI project you are working on.

A data security policy typically includes two categories: people and technology. The people elements of the policy can cover segments such as acceptable use, security incident reporting, passwords, social networking, or emailing. The technology elements of the policy can include encryption, access management, system security, vulnerability scans, backup, and recovery or mobile device management.

See an example of a data security policy template

6. Anonymize or pseudonymize data

Regarding personally identifiable data, we only collect as much as needed for a specific case. And when possible, we run algorithms for anonymizing or pseudonymizing this data before processing it in our systems. For example, we have models available for replacing faces in images and can also anonymize identifiable information in the text before processing it with our users.

This is data category specific as not all algorithms are applicable for all use cases.

While pseudonymization and anonymization both refer to hiding personally identifiable data, their methods differ. Pseudonymization masks the data to the extent that the person can no longer be identified without using additional information. Personal data is replaced with other, non-identifiable data, and additional information is needed to recreate the original data. Meanwhile, anonymization masks the data so it can no longer be identified. In this case, no additional information will recreate the original data. Anonymization comes with a degree of complexity as it eliminates the connection between data and the individual. For this reason, anonymization is more commonly used for statistical or research purposes.

For example, even if some identifying information was removed, you can still recognize a person from a voice recording if you have a database of voice recordings to compare it to. But if the voice recording was anonymized (transformed, distorted), it would no longer be possible to identify the person.

7. Prepare a DPA to comply with GDPR

If we handle data that comes from the EU and contains PII or biometric data, we sign a data processing agreement (DPA) or a controller agreement with the client to comply with the general data protection regulation (GDPR). These agreements specify all the necessary guidelines and procedures regarding data handling.

See a data processing agreement template.

8. Sign the contract to deliver data

The final step. After the case is evaluated, solutions are suggested, risks assessed, and all the relevant data agreements signed, it is time to sign the contract and then collect and deliver the training data.

Need data? We are here for you.

When developing your AI project, you may be faced with challenges that are unique to your business. We are prepared for that. StageZero is fully equipped to provide data of various types and languages.

Once we're sure we can collect the right data for you, we can engage our global crowd for rapid scaling. We're 100% GDPR-compliant, and the security of your data is ensured when transmitted, stored, and processed.

Learn more about how we can help you with training data and regulatory compliance.

How to develop GDPR-compliant AI

AI development keeps growing in popularity and sophistication, requiring more and more data, often sensitive data, to function and predict. This brings AI straight under the scope of the General Data Protection Regulation (GDPR). Since its implementation, the European privacy and security law has reshaped how companies handle data. Be it concerning cookie policy or AI development, companies need to follow the principles of the GDPR; otherwise, they risk hefty fines.

Why GDPR applies to AI development

The GDPR was created and enforced by the European Union (EU) to protect personal data privacy. Provided they target or collect data related to individuals in the EU, organizations worldwide are obligated to comply with the GDPR.

Each EU member state supervises GDPR compliance through independent data protection authorities. The GDPR came into effect in 2018, and fines up to €746 million were already handed to companies in multiple industries. For example, Clearview AI paid a €20 million fine for failing to process personal biometric and geolocation data lawfully. The company also breached several fundamental principles of the GDPR, such as transparency, purpose limitation, and storage limitation (more on the principles later).

So while the GDPR does not explicitly mention AI, companies developing AI models must comply with the regulation: AI development uses data, and GDPR protects that data.

How GDPR impacts AI development

Any AI development data collected within the EU falls under the protection of the GDPR, regardless of whether your company is based in or outside of the EU. Regulatory compliance is mandatory as long as it is related to EU residents.

What is important to note, personal data collected before the GDPR lacks such legal compliance and cannot be used. The legal ground for collection must be presented to the person the data is collected from. In other words, the person needs to know for what purpose you are collecting the data and give their legal consent in the form of a contract or other means.

The GDPR is based on principles that must be followed when developing technologies involving data. Most of these principles are also highly relevant to AI development. The principles provide data protection guidelines for organizations looking to develop AI models and collect necessary training data.

Purpose limitation

Purpose limitation represents limiting the repurposing of personal data. For example, data collected for a customer service use case can be reused to send tailored marketing messages, but this has to be clearly indicated. Establishing whether the data reuse is legitimate will depend on whether a new purpose is compatible or incompatible with the original intent to collect data.

To avoid breaching the principle of purpose limitation, specify the purpose of data collection and usage in your data privacy policy. This should cover all possible use cases upon collection.

Data minimization

At first glance, data minimization might not seem compatible with AI, as its models need lots of data to learn. But data minimization simply encourages a judicious approach to data collection, and it is about focusing on the quality of datasets rather than large volumes of data.

According to the GDPR, companies should not process any more personal data than is needed to reach their goals. To comply with the principle of data minimization, in your privacy policy indicate what PII data was collected (if any) and specify how long data will be stored and under what conditions.

Automated decision-making

As the term suggests, automated decision-making is done without human involvement. Automated decisions can be based on factual, digitally created, or inferred training data.

Concerning AI, automated decision-making is the most commonly discussed GDPR principle. That is because AI implies automation by nature, and under the GDPR, individuals have the right not to be subjected to a decision solely based on automated processing.

When collecting data, companies are obligated to provide information about the logic, significance, and consequences of the decision to the persons whose data is being collected. This information has to be presented in clear and plain language.

Automated decision-making is closely related to profiling, which brings us to the next principle.

Profiling

Improving technological AI capabilities increased the opportunities for profiling dramatically. This principle refers to an automated data processing to assess the data subject's personal aspects. Such aspects can be utilized, for example, to analyze a person’s economic situation or work performance. The data is then used to develop comprehensive user profiles, for example, for creating tailored advertising.

Some uses of profiling can lead to unwanted results. For example, Amazon had to scrap its AI recruitment tool after it proved to show bias against women.

According to the GDPR, it is prohibited to subject someone to any decision based solely on automated processing, where the decision has legal or another severe effect. Profiling can never be executed based on race, religion, or health data unless explicit consent is given or it is in the public’s interest.

Fairness

Following the fairness principle, companies must not process data in an undisclosed or ill-intentioned way. AI models must not use data to generate adverse outcomes for those whose data was processed.

While it is rare for a company to knowingly employ unfair AI data practices, unfair data use can often occur unintentionally. For example, Twitter’s image cropping algorithm was deemed racist after users noticed that the feature automatically focused on white faces over black ones. Biased data breaches the principle of fairness and produces flawed AI models. However, you can avoid it in your AI development by using diverse training datasets.

Transparency

According to the transparency principle, individuals must be fully aware that an AI system will process their data. Subjects must be well informed on data processing purposes and understand how the applied AI algorithm has come to a decision concerning them.

To ensure transparency when developing your AI model, you should provide the information above in your company's privacy policy or elsewhere on the website.

Accountability

Whether because of lacking setup or biased data, AI-based decisions always have the potential for negative outcomes. Proactive measures can vastly minimize these outcomes. Following the accountability principle of the GDPR, companies must implement strategies and procedures to predict or mitigate data privacy risks.

Create a Data Protection Impact Assessment (DPIA) to show relevant authorities that you have thought about risks and mitigation. The policy will help reduce the probability of getting a fine if something goes wrong.

How to ensure GDPR compliance when developing your AI model

If you follow the GDPR principles listed above, you are already on the right track to ensuring compliance. Additionally, there are a few specific methods you can apply in your AI development:

- Reduce the need for training data by applying methods such as federated learning.
- Uphold data protection without reducing the primary dataset with differential privacy or homomorphic encryption.
- Avoid the "black box" issue with methods such as explainable AI (XAI).

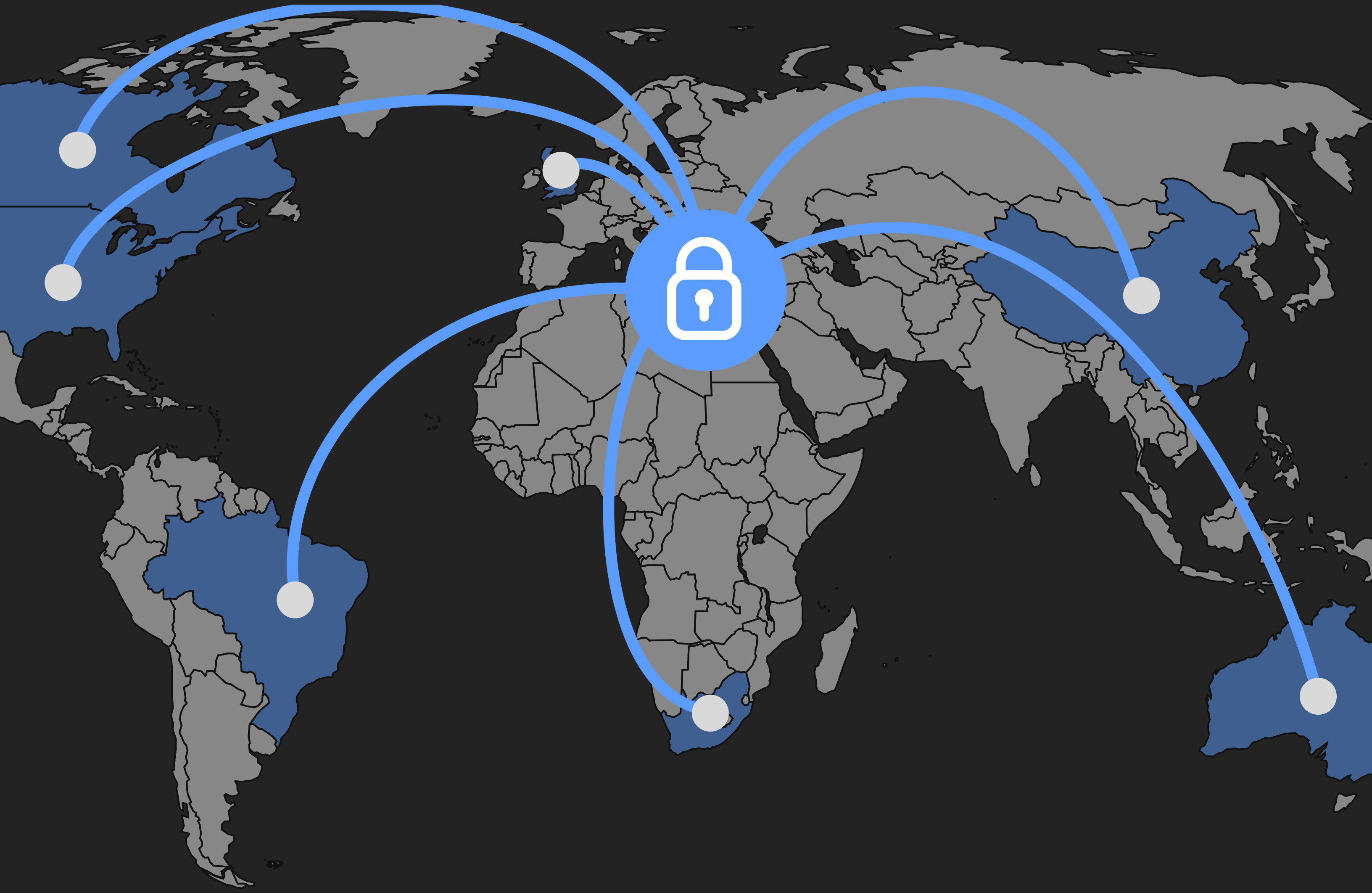
But by far, the best way to avoid paying a sizable fine is to focus on the data. Use diverse, carefully selected datasets to prevent bias. For example, remove racial components in use cases where that could cause biased predictions. And whenever possible, train AI models on anonymized data.

Whether using or collecting your own data or collaborating with a data vendor, ensure all privacy-related regulations, including the GDPR, are followed through.

Need help? Reach out to StageZero if you want to know more about data collection and necessary regulatory compliance.

AI and regional data privacy laws: key aspects and comparison

In recent years, a number of data privacy acts around the globe have come into effect to fight data exploitation. While not explicitly targeting AI, these regulations protect data, automatically pulling AI development under their radar. Each privacy act comes with its set of requirements and financial penalties; find out how they differ and what you can do to ensure regulatory compliance when developing your next AI project.



Why data privacy regulations are important

To produce intelligent, well-informed decision-making, AI models must be trained using vast volumes of diverse data from myriad sources. At the same time, dealing with large unstructured datasets from multiple sources comes with a privacy risk and possible legal repercussions.

More and more data privacy acts around the globe are introduced to safeguard people from data breaches and re-identification. These regulations protect individuals, help reduce cybercrime, and push companies using training data for machine learning to do better.

Data protection laws should be more than something businesses must overcome to avoid fines. Following the guidelines underlined in the privacy regulations, companies can ensure compliance but also maintain a more streamlined data management framework because data privacy acts are based on the best practices for data processing and security.

Data privacy acts around the globe

Data privacy regulations vary across different continents, countries, or even states. Still, all typically address the same issues: what data needs protection, how to protect it, and what happens if you don't. Below we look at some of the most significant data privacy acts worldwide and explore their relation to AI development.

The General Data Protection Regulation (GDPR)

The GDPR is probably the best-known data privacy act globally. The regulation came into effect in 2018. Since then, many of the subsequent privacy regulations in other parts of the world were, in fact, heavily influenced by the GDPR. According to the regulation, companies are obligated to secure user consent to use their data. This also applies to historical data, which caused a headache for many companies hoping to use their internally collected data in AI development.

Who it applies to: this regulation protects the data privacy of EU residents and applies to any organization that processes that data, regardless of where the organization is located.

How it connects to AI: the GDPR is based on a set of principles that must be adhered to when handling personal data: purpose limitation, data minimization, automated decision-making, profiling, transparency, fairness, and accountability.

While most of these principles apply to AI development, automated decision-making and profiling are the two you need to be especially aware of. According to these principles, companies are required to provide information about the logic, significance, and consequences of automated decisions to the people whose data is being collected.

It is also prohibited to subject someone to any decision based solely on automated processing, where the decision has legal or other severe consequences. A decision will not be considered solely automated if a human evaluates the result of an automated decision before applying it to the affected person.

Failing to comply: organizations can be fined up to 4% of their global annual revenue or €20 million, depending on whichever is higher.

The California Consumer Privacy Act (CCPA)

The CCPA came into effect in 2020. Just like the GDPR, the Californian data privacy act is created to prohibit the exposure of personal data without users' knowledge and to protect consumer rights regarding their data. The regulation differs from the GDPR in that you don't need to obtain prior consent to simply collect and process personal data.

Who it applies to: organizations that collect and use personal data about citizens of California. While it is "only" a state law, California is the most populous state in the United States, so, in reality, the CCPA applies to any U.S. or international company that does not want to exclude the data of a large part of the American population. Plus, there is the fact that if California was a sovereign country, it would rank as the world's fifth largest economy.

How it connects to AI: the CCPA doesn't specifically address AI or its use. Yet, the regulation defines user rights, such as the right to know, the right to deletion, the right to opt-out, and the right to non-discrimination. These rights can be interpreted as directly applicable to data management in AI. Companies need to clearly and transparently disclose AI usage in their purposes for collecting and processing data and remove user data when requested.

Failing to comply: a fine of up to $7,500 for each intentional violation and $2,500 for each unintentional violation. The penalties are defined per user so; for example, if your violation is defined as intentional and affects 100 users, you could end up paying $750,000. Personal claims by an affected user can also be made for up to $750 per violation.

The UK General Data Protection Regulation (UK GDPR)

The UK GDPR is the United Kingdom's data privacy law. It is essentially the same as the GDPR, with changes made to accommodate domestic UK law areas such as national security, intelligence services, and immigration. The regulation was drafted as a result of the UK leaving the EU, relates to the UK'S Data Protection Act 2018, and came into effect in 2021.

Who it applies to: the data privacy law governs the processing of personal data from individuals inside the UK. Any entity inside or outside the UK has to be compliant if it collects or uses data from individuals in the UK.

How it connects to AI: the UK Information Commissioner's Office, which oversees the application of the UK GDPR, has released its guidance on AI. Similar to the European GDPR, the document presents a compliance framework by introducing a few principles.

The principles are grouped into four parts: accountability and governance implications; lawfulness, fairness, and transparency; assessing security and data minimization, and ensuring data subject rights. Based on these principles, companies are recommended to have DPIAs[1] , keep transparency and clarity when documenting their data processes, implement effective risk management practices, and set up systems to effectively respond to and comply with data subject rights requests.

Failing to comply: a maximum fine of £17.5 million or 4% of annual global turnover - whichever is greater.

The Personal Information Protection Law of the People's Republic of China (PIPL)

The PIPL came into effect in 2021. It regulates the collection, use, and disclosure of personal data and is partly based on the European GDPR. The regulation mandates companies to get consent from individuals before collecting their data. According to the PIPL, individuals have the right to know what data is being collected about them and how it will be used.

Who it applies to: all organizations that process any personal data originating in China.

How it connects to AI: similar to other privacy acts, the PIPL is addressing automated decision-making. It regulates the use of algorithms and other automated systems that can discriminate against certain individuals. Consent is required in most circumstances. Among other guidelines, the PIPL requires companies to conduct audits to assess data security risks and implement safeguards.

Failing to comply: penalties of up to RMB 50 million, 5% of a company’s annual revenue, and seizure of all illegal gains.

Other data privacy regulations

Aside from those discussed, data privacy laws similar to the GDPR have been introduced elsewhere in the world: Canada's PIPEDA, Brazil's LGPD, Australia's Privacy Act and Consumer Data Right (CDR), and South Africa's POPIA, among others.

If you are GDPR-compliant, are you compliant with other regulations by default?

Since the GDPR is the basis for many other data privacy regulations, the natural question is whether it is enough to follow the principles of the GDPR. However, while most data privacy acts that came into effect in recent years share lots of similarities, you should continually assess all regulations applicable to your specific case.

Regulatory compliance in AI development: where to begin

New data protection regulations are introduced every year. As you look into your options to source training data for our AI project, keeping track of all you need to comply with might be challenging. We recommend beginning with:

Evaluate what data you have/need

The type of data you store or need will determine which data protection regulations you are required to comply with.

Prepare a data compliance plan and regularly reassess

Have a detailed compliance plan for how to achieve regulatory compliance and address any unexpected risks. We recommend a regulatory compliance checklist[1] . As time goes on and new data privacy acts emerge or current ones change, regularly reassess your data compliance status. Perform regular data assessments that help identify areas for improvement.

Seek legal advice

When in doubt, seek legal advice. Reach out to lawyers who specialize in regulatory compliance.

How StageZero can help with your AI development

Compliance with data privacy regulations starts with quality datasets. And with the right data vendor, this part is less of a burden.

We specialize in sourcing diverse, high-quality datasets of various types and multiple languages. The right datasets lead to the right results and ease the process of meeting legal requirements all at once.

We are here to help you satisfy your machine learning data needs and share our tips regarding regulatory compliance. Simply reach out and tell us about your AI project.

🌐 stagezero.ai

in StageZero Technologies