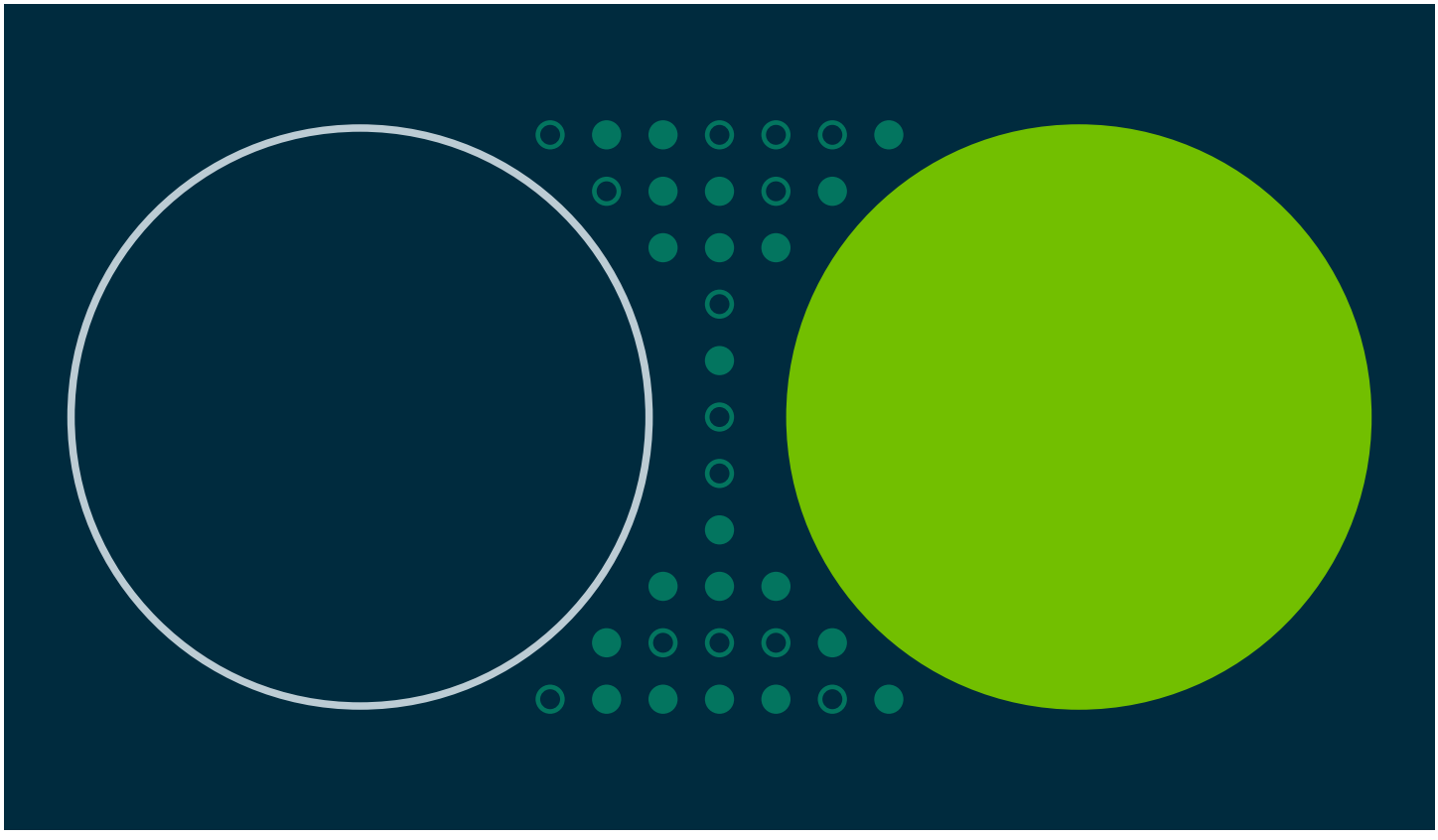
Collibra

Whitepaper

Data ethics in a digital age

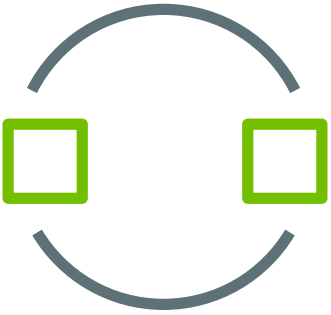
What it is, why it's critical, and what you can do to drive adoption at your organizations

January 2023



Introduction

Data ethics. It's a concept data professionals find themselves facing more and more, especially in the age of GDPR and the ascent of AI. It's relevance today points to the power of data in our data-driven digital economy. And with great power comes great responsibility.



Today's data professionals are capturing more data than ever before. They're investing in data infrastructure. They're leveraging powerful AI models. Getting more value out of data is at the top of every CEO's list of strategic initiatives. But data by itself does nothing. It requires people to make critical decisions about how to leverage it. As organizations around the world are challenged to develop AI models that are bias-free, the significance of data ethics has taken center stage.

But data ethics is about much more than strategies to eliminate bias in AI training models.

In some ways, the focus on data ethics makes an odd marriage of two very different ideas with very different histories. On the one hand, the idea of ethics is as old as Aristotle who wrote his *Nicomachean Ethics* around 340 BCE. Data — in the sense that we use it in our computer age — has only been around since the 1950s.

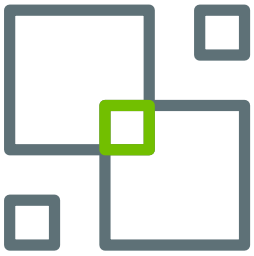
Data is informational. It's factual. It's numeric and quantifiable. As data professionals, we are familiar, even passionate, about the precision of our exacting science. On the other hand, ethics is seemingly much more open to interpretation. We say that ethics is about doing the right thing. But what is the right thing? And who says what's right and what's wrong?

Ultimately, each organization chooses its own data ethics framework. And every organization understands the risks of falling afoul of cultural norms about what is right and wrong — the costs to reputation, to revenue, to staff retention.

This whitepaper will shine a light on data ethics. We'll cover how data ethics touches each stage of the data life cycle. And we'll help you design and drive adoption of a framework for making ethical decisions about your enterprise data.

"Data ethics is defined as 'the moral obligations of gathering, protecting, and using personally identifiable information and how it affects individuals.' "

Source: Catherine Cote, 5 Principles of Data Ethics for Business, HBS Online.



Why data ethics is essential in our digital era

Today, the intersection of data and ethics plays out in our personal lives and in the life of our society because of the power of data. From the living room to the boardroom to the hospital bed, data drives decision-making. During the Covid-19 pandemic, we saw governments prioritizing vaccine shipments based on positivity rates. A recent U.S. study showed people check their phones 344 times and receive 46 push notifications per day. In so many ways, data is the currency that makes the world go around. But it also presents an opportunity for unscrupulous marketers and an attack vector for cyber-criminals.

The scale and velocity of big data pose a serious concern as many traditional privacy processes cannot protect sensitive data, which has led to an exponential increase in cybercrime and data leaks.

Source: Dominique Daly, The Ethics of Big Data.

A recent U.S. study showed people check their phones 344 times and receive 46 push notifications per day.

Data professionals like us are at the center of our modern challenges to unlock more value out of data while valuing moral considerations around fairness, ownership, privacy, and much more.

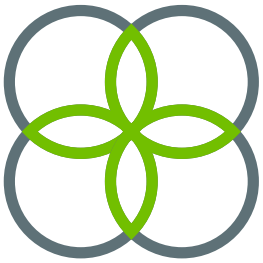
It's a challenge that encompasses much more than AI and ML. If you're collecting data about people, then there are ethical choices to make about your data. In this way, the ethical enterprise leverages an existing data life cycle — and applies data ethics to it.

The ethical data life cycle

The data life cycle describes the stages that data moves through from creation to deletion. It is a map of the flow of data through an organization. It describes how data is collected, stored, used and eventually destroyed.

When we talk about an ethical data lifecycle, we're talking about the process of ensuring ethical considerations are applied to data as it moves through every stage of the data life cycle.

Data ethics is the set of principles that govern how data should be collected, stored, analyzed and used.



There are a range of models representing the data life cycle. For this whitepaper, we'll use the following four-pillar model:

1. Collection
2. Storage
3. Usage
4. Retention

"Even if you don't directly work with your organization's data team or projects, understanding the data life cycle can empower you to communicate more effectively with those who do. It can also provide insights that allow you to conceive of potential projects or initiatives."

Source: Tim Sobierski, 8 Steps in the Data Lifecycle, HBS Online.

The following sections discuss each stage of the ethical data lifecycle:

1 Collection: What are we collecting? Why are we collecting it?

An ethical approach to data starts with the first stage of the data life cycle: collection.

Data collection is the process of gathering data for decision-making, analytics, and other business-critical applications. It includes all activities involved in obtaining data from a variety of sources, including IoT ecosystems, operational systems, financial transactions, business partnerships, and much more.

- A data collection process can involve a range of steps, including:
- Identifying what you want to measure and how often
- Deciding who will collect it
- Selecting a method(s) appropriate for its purpose
- Piloting the collection

- Recruiting respondents (if applicable)
- Securing access permission (if applicable)
- Collecting responses using agreed upon approaches
- Coding responses according to specified rules — whether manually by hand or electronically using machine learning algorithms
- Entering data into a database securely

These steps may vary but usually involve ensuring compliance with relevant laws. So compliance precedes all other considerations.

When talking about data collection, it's important to understand that this is the first step in your data lifecycle — and the first place where you can start asking ethical questions about how you collect data, including:

- What kind of data are we collecting?
- Where does this data come from?
- How clearly have we expressed the purpose?

Transparency: Essential to data collection

A crucial consideration at the Collection stage is transparency. You should strive to be as transparent about the purpose of your data collection and whose data it is that is being collected. Ensuring that your purpose for data collection is crystal clear is essential.

Each one of your customers or end-users must understand that, for example, by accepting cookies and agreeing to your terms on your website, you may share their behavioral and demographic data.

At the end of the day, you are contributing to the make-up of your data ecosystem and to all the data that becomes the basis for policy making. Make sure you include this fundamental consideration when applying data ethics to your data life cycle.

2 Storage: Where are we storing and how?

Data storage is the second step in the data lifecycle. From an ethical perspective, there are several factors to consider when determining where and how to store your data, including:

- Who needs access to this data?
- What are the geographic/legal requirements?
- What is the sensitivity of this data?
- What are your business and security needs?

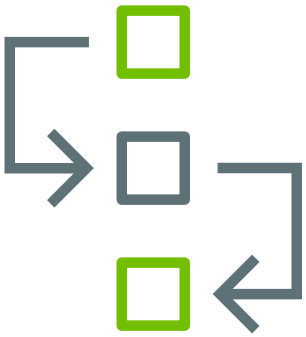
By classifying the data from the get-go, you can address a lot of problems down the line.

For example, data sensitivity drives the need for security as well as access considerations. By defining your data sensitivity parameters, you can better understand your organization's security requirements.

However, the need to continuously drive effective data security during the Storage phase highlights the demand for well-defined, documented, and easily accessible data classification and data security policies. No one wants to review an entire policy when you're working on a single data set.

An ethical data life cycle thrives when there are design processes that make it easy to do the right thing. You can do this by breaking down your comprehensive policy into smaller, logical chunks. Then your data can be appropriately tagged with the relevant policy sections.

And by establishing a data catalog, your organization can keep track of where data lives, no matter where it's stored.



3 Usage: How are we using it?

Usage is the third stage in the data life cycle. It is critical because it's not just about the data itself, but also about who is using it and how they're using it.

In this phase, ethical considerations include:

- How are we using data and for what purpose?
- Is our data usage consistent with what we declared during collection?

An analyst or a data scientist is far removed from the data collection stage and cannot be expected to be aware of the purpose that was shared with the end-user.

A robust data ethics framework can address this gap by ensuring the data comes with a data-sharing agreement that states acceptable data usage. The key element is to be clear, but not necessarily exhaustive.

So what do we do when the use case is not completely aligned with the acceptable uses? When data ownership is established and is available easily, it becomes a matter of clarification. It can then lead to expansion of scope for data usage or redirection to another data source. You can make it easy to do the right thing!

Questions to ask to limit bias in AI models:

When it comes to decisions, how will this data product contribute?

- Assist a human decision-maker
- Replace human decisions that require judgment or discretion

Are the impacts resulting from the decision reversible?

- Irreversible
- Difficult to reverse
- Reversible

How long will impacts from the decision last?

- Less than a week
- Less than a month
- Less than a year
- Years

Are all decision points well documented and accessible to all stakeholders?

- Yes or no

The challenge with bias

In the usage phase of the data lifecycle, there is also a greater focus on bias, especially unconscious bias in AI models.

This is an area that gets a lot of attention. At Collibra, one of the ways we are beginning to address bias is by asking thought-provoking questions early in the design phase.

Asking serious questions about bias in your data usage policy generally elevates awareness of the issues at hand and the need for a deeper and wider data set.

Transparency: Essential for data usage

Transparency is not only essential in the Data Collection phase, it's critical in the Usage phase, too.

In the Usage phase, transparency implies you are not only aligning with the purpose you established in the Collection stage, but you can explain how you align.

However, when you create blackbox machine learning models, you lack this ability. Therefore, the interpretability and/or explainability of a model is a key consideration at the design phase for data scientists.

By making the purpose established at the Collection phase available to analysts and scientists, you can open avenues for capturing usage evolution; you can design a feedback loop into the creation of purpose language, such as privacy usage agreements.

As you build out your capabilities, it's also important to have a team with diverse experience and perspectives.

4 Retention: How long are we keeping it?

Retention (and deletion) is the final phase of the data lifecycle. It is important for every organization to have well-established data retention policies.

Think about it: You've created an ethical framework for the first three stages of your data life cycle — data collection, storage and usage. But if you haven't established parameters around how long you retain data, you're creating unnecessary risk.

Here are some questions to ask yourself about the Retention phase to ensure the health of your data lifecycle:

- Is your organization storing data that is never used?
- Are you tracking data usage and using that analysis as a feedback loop to inform your approach to the earlier stages of the data life cycle?
- For example, do you really need to store data that will only be used for real time processing?

Some of these policies may be governed by your company's compliance with industry regulations, but others will be based on its own internal rules about how long it retains certain types of data. You should also think about how your organization uses or expects to use the data in question before establishing any policies around retention and deletion.

As consumers become more privacy-conscious, there's been an uptick in the use of the right to erasure. It then becomes increasingly important to be able to track data and delete easily.

When you're looking for a solution for systematic retention, it's important to have a clear understanding of these different types of requirements in order to find the right solution.

So how do you anchor your data strategy to an ethical framework that works for your organization?

Start with what you believe in

When you're ready to begin your data ethics journey, the first place to start is looking at your company values. (If you haven't yet established and communicated your values, this is a great opportunity to do so.)

At Collibra, our corporate values speak to the kind of organization we want to be. Two of our values — 'Open, direct and kind' and 'Our work matters' — have direct relevance for how we make decisions about data. We use these values to guide our roadmap.

Key terms in data science

Interpretability: The ability to present in understandable terms to a human how a prediction was derived by inspecting the model itself. In other words, interpretability refers to the resulting prediction being readily discernible directly from the inputs, by a human. This is highly desirable.

Explainability: A set of techniques, often applied to black-box models, to explain a prediction.

People are talking about data ethics

At the Collibra Data Citizens '22, data ethics was top of mind. In fact, our 'Data Ethics' session was one of the most popular of the event.

Go to Data Citizens

Collibra | Our values

Open, direct, and kind
Collibra is a place where we speak openly about what's on our minds. We care about each other so we are thoughtful about what we say, when we say it and how it's said.

Our work matters
We are passionate and talented people who take ownership and get things done. Our ideas and actions will transform our company, our industry and a lot more.

The 3 essentials of a data ethics initiative

- Be transparent about your data practices
- Be prescriptive with your policies and guidance
- Be values-driven in your data ethics strategy

Consider the impact

While values are great guidance, the impact of data is where it gets real. It is important to ask those thought provoking questions about everything you do with data.

Be good to your data

Your values will direct your data ethics journey and determine how you focus your efforts. As a forward-thinking organization, Collibra is focused on strengthening our data intelligence practice by formalizing our organization-wide position on data and our decisions about data.

Conclusion

Data-driven enterprises need an ethical framework to guide data management. Informed and inspired by your values, your organization may be on the road to standing up a data ethics initiative.

When you do, your next step will be to make sure you have trusted data, and that usually requires three critical capabilities:

- The ability to locate your data easily
- The workflows and processes to collaborate across your business
- The existence of a unified view of all your data assets

The practical application of data ethics can ensure your enterprise data is used ethically and responsibly throughout its entire lifecycle: from taking an action based on data insights, through making sure that all of your employees are adhering to policies related to privacy and security (like GDPR), right up until when you discard or delete that information.

It is important to remember not to boil the ocean, you don't have to do everything at the same time. Keeping it simple will help you succeed in your implementation. Set clear expectations, ask thought provoking questions and provide alternatives. This will make it easy to do the right thing!

Learn more about how Collibra can help you on your data ethics journey.

