Sonasoft

How to solve
AI Debt

Executive summary

AI is driving the 4th industrial revolution. As with any revolution, if you ignore it your company will get left behind as an anachronism. AI drives innovative ways to do business. It enables firms to offer 10-minute grocery deliveries. It powers virtual assistants that transform our home lives.

Unfortunately, despite its importance, most AI initiatives fail to deliver the expected transformation. AI debt hamstrings businesses of all sizes and prevents them from realizing the benefits that AI can deliver.

What is AI debt?

AI debt is a measure of how far behind you are with your AI adoption strategy. In many ways, it is similar to the technical debt that every CTO knows well. The difference is, AI debt is often intrinsic. There are three main forms of AI debt you should know about.

- **Data AI debt.** This reflects the difficulty in getting the right data in the right place at the right time. Without data, AI simply cannot exist. Yet, most of your data is spread across heterogeneous systems, often in legacy formats that make it hard to access. You face the challenge of combining all this data into one cloud-based data platform.
- **Technical AI debt.** This is the closest analog to classic technical debt. It reflects the difficulty in productizing AI solutions and launching them in your production environment. This is due to the way that AI projects are often developed as one-offs with custom code that cannot be reused or simplified.
- **Human AI debt.** The final source of AI debt is your team. AI is still a relatively new concept. As a result, specialist AI-focused roles are still emerging. These include such things as MLOps engineers and data engineers. Currently, very few people have the skills needed to do these roles. This makes it very hard (and expensive) to build a strong AI team.

All these are equally damaging and all are hard to eliminate.

How Sonasoft helps

At Sonasoft we view AI as an end-to-end problem. We work with you to help deliver successful AI initiatives starting from the data you already hold. We then develop a solution using SAIBRE, our AI Ecosystem. This can be easily deployed to production where our smart monitoring system helps ensure it keeps running correctly.

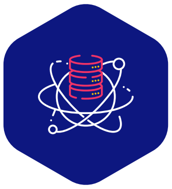
Why is AI critical for your business?

We are about to see a paradigm shift in the way we do business. Up till now, AI has been a “nice to have” feature for most businesses. But over the next few years it will become as critical to your business as the Internet itself. This transformation has been accelerated by the global pandemic, which has seen traditional approaches to doing business being challenged everywhere. Once, every tech team had to share an office with all interactions happening face-to-face. Now, we see that modern technology allows us to blur the distinction between remote and in-person working. Shopping used to be about bricks-and-mortar stores. Nowadays even your last minute purchases can be delivered in minutes by the likes of Door Dash and Gorillas.

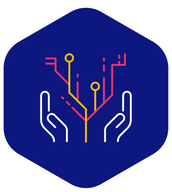
Underpinning a lot of these changes are AI algorithms. British online supermarket, Ocado, uses autonomous robots to pick and pack customer orders in their huge central warehouse. AI is used to plan optimal delivery routes that maximize the number of deliveries while minimizing the miles traveled. Most customer service teams have outsourced their first contacts to chatbots that are growing ever-more sophisticated. Virtual call centers mean that call handlers are able to work from home without any loss of productivity. Intelligent call redirection ensures customer calls are directed to the person best able to help them. AI is also helping companies keep their network perimeter secure even while most of their workforce is working from home.

What is AI debt?

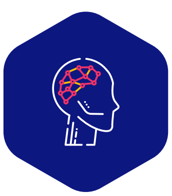
As AI grows in importance, we are starting to see a new phenomenon we can call AI debt. This can be thought of as being akin to technical debt. All CTOs and VPs of engineering will be familiar with technical debt but few people have yet to address AI debt. AI debt comes in three distinct forms. Each of these will have a different level of impact but to solve AI debt you have to tackle all three.



Data AI debt. AI solutions rely on data at every stage. Before starting, you have to collect all the data in one place and make sure it is in a readable format. This can be a real hurdle, with data spread across multiple systems in many legacy formats. As a result, you often need to undertake significant data engineering and cleaning. During model selection and training, the data has to be accessible over a fast connection, else your process will grind to a halt. Once you deploy the model to production, it needs a constant live feed of the latest data, both internal and external.



Technical AI debt. Traditionally, technical debt arises when your software relies on code that is brittle, poorly designed, or unable to scale. This is a fact of life in any rapidly scaling business where functionality outweighs performance while you build market share. In AI, technical debt has similar impacts: models are brittle, they are not designed to integrate with your systems, and they don't work at the scale needed. Moreover, unlike classical technical debt, technical AI debt comes back whenever you build a new model because your team likely starts from scratch each time, rather than creating reusable components.



Human AI debt. AI development is still in its infancy compared with other forms of software. As a result, the roles within the team are still quite poorly defined and tend to be siloed. Moreover, there is a global shortage of really good talent—many people are claiming to be data scientists or ML engineers, but few of them are all that competent. Even the competent ones are usually coming from an academic background with little experience of the business world. Then there are the problems across your wider team. Even if you manage to assemble a top-notch MLOps team, people in the rest of the company may not understand AI. That means they are unable to provide clear problem statements or assessments to the AI team.

How does AI debt impact your business?

AI debt can be a problem for almost any company, regardless of where they are on the AI journey. Companies with no AI at all face an uphill struggle against competitors who are already leveraging AI. Companies who have adopted an AI strategy often find that their strategy fails due to AI debt. Finally, companies who have already deployed AI solutions will find that AI debt has a habit of coming back as their solutions age.

Lack of an AI vision. Over the next couple of years, companies without an AI vision are going to suffer more and more from their growing AI debt. They will lose customers to competitors, suffer from reduced productivity, and will eventually become sidelined.

Failure to deliver your AI transformation. As many as 87% of AI projects never make it to production[1]. That's a staggering number of failed projects and equates to billions in wasted investments. In most cases, the root cause of the failures is AI debt as we will see later.

Failure to keep your AI up to date. At the heart of every AI solution lies one or more machine learning models. These models rapidly become stale so, over time your AI debt starts to come back. The upshot is that you lose the competitive edge again.

AI debt impacts almost every aspect of your business. But, if you'll pardon the pun, the bottom line is it's your bottom line that will suffer! The three biggest problems you face are as follows.

Failure to deliver savings

Typically, companies turn to AI because it promises significant cost savings through improvements in efficiency, reduction in waste, etc. Often, companies will sell you AI solutions by claiming they can make huge cost savings: "halve your costs overnight", "boost your team's productivity by 100%", etc. In fact, successful AI projects often only deliver improvements of 10-20%. But with most companies looking at profit margins of 1-2%, that can easily lead to a 10-fold boost in profitability.

Loss of opportunities

AI is a truly disruptive technology. Failure to adopt it will leave your company lagging far behind your competitors. Put simply, AI will allow them to undercut you at every

1 https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/

stage. Amazon became the world's richest retailer because of the power of AI. That allows them to deliver true dynamic pricing and intelligent recommendations, driving up their profits and edging competitors out of the market. Likewise, Google became so dominant because it leverages AI in every part of its business. If you don't adopt AI, you will soon find yourself lagging far behind your rivals.

Stale models

After delivering a major project it can be tempting to sit back on your laurels and think all your problems are solved. But any product manager will tell you that's not true. However good your product is now, you can still improve it and deliver something better to the end users. AI is the same but it also has an additional problem that doesn't affect most other products. AI models can (and do) go stale. Specifically, over time what you get out of your model will become less and less accurate. This is especially true for any form of forecasting model. This form of AI debt can be especially damaging because often you will have come to rely on the model being accurate. When it proves wrong, you may face a heavy financial hit.

Why is it so hard to get rid of AI debt?

AI debt is remarkably stubborn and hard to shift. Solving it needs buy-in from all parts of the company from C-suite down. A 2021 article from McKinsey points out that the CEO plays an essential role. The authors talk about the importance of MLOps as a philosophy, and go on to say:

> "As a result, CEOs play a critical role in three key areas: setting aspirations, facilitating shared goals and accountability, and investing in talent." [2]

MLOps is becoming a bit of a buzzword in the AI industry. It is loosely based on applying the principles of DevOps to the creation and deployment of machine learning solutions. It's hard to find any clear definition of the term. The one chosen by McKinsey is quite broad:

2 https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/scaling-ai-like-a-tech-native-the-ceos-role

“MLOps establishes key practices across the application life cycle that increase productivity, speed, and reliability and reduce risk. “

Essentially, MLOps involves taking the best practices from software engineering and applying them to AI. Thus, the focus is on creating reusable components and techniques, streamlining the development process, designing solutions for deployability, and ensuring you maintain the solution in production. Like DevOps, MLOps is more of a philosophy than a rigid set of guidelines. So, how can you apply this philosophy in practice?

Always start with the data

Without data there is no machine learning or AI. So a successful AI project is predicated on having access to sufficient high-quality data. Chances are, your company has data scattered across every business unit. It’s not uncommon to find one company relying on a dozen different systems from Hubspot and Salesforce to MySQL and Amazon S3. Two problems often arise. Firstly, each team developing AI solutions typically starts from the raw data you hold. Each team will go through the process of importing this data, exploring it for interesting features, and cleaning it up. Secondly, because this process is much more boring than building ML models, many teams will fail to put in the required effort to do it properly.

What issues might you face?

This traditional approach to data discovery and exploration creates clear issues that go on to impact the success of the resulting AI solution.

- **Incomplete data.** Often, your team will either lack the expertise, time, or resources to collect all the possible data that might help their model. Instead, they will focus on a subset that’s easier to get hold of or is directly under the control of the business unit they are developing a solution for.
- **Unexpected bias.** It is all too easy to introduce bias into your data. ML models aren’t infinitely powerful and often there are resource constraints. As a result, your team undertakes feature engineering to try and clean and simplify the

data. But this almost always adds some bias. The trick is to know that there is bias and to compensate for it in the next stage.

- **Single-use datasets.** Most teams only focus on the immediate problem they are working on. Thus, during their data exploration and feature engineering, they often end up creating a dataset that is only suitable for one problem. That means the next time you have to go through the entire process again. Not great, since this can take a couple of months!

How can you solve this?

The MLOps ethos is to make reusable components to streamline the process. One of the easiest wins is to focus on creating reusable datasets. For instance, if your team is working on a project to create a lead scoring system for the sales team, they will need all data on marketing leads and customer deals. If treated right, this dataset can also be used for other projects, such as a deal recommender engine to help your sales team convert their leads.

Create deployable solutions

Often, your AI team will lack experience in software engineering. Typically, they will come from a data science background. That means they are great at crunching data and building ML models, but less good at creating the supporting code that allows those models to be deployed. One of the most frequent reasons an AI transformation fails is because the solution can't be easily deployed and run in your production systems.

What issues might you face?

Clearly, if your new AI solution cannot be deployed at all then it's doomed from the start. However, very few companies will allow this to happen. Instead, they will divert resources from the DevOps and systems teams into solving the problem. They will try to shoehorn the AI solution into the production system and hope for the best. There are a couple of frequent problems we see.

- **Incompatible architectures.** Modern systems tend to follow one of three distinct architectures. The classic is the client-server model, typically nowadays using a virtual server in the cloud. The second is a microservices architecture, which allows rapid scale-up and great flexibility. The third is a containerised architecture, where each application comes as a bundled set of services within a container. The problem comes when you try to connect different architectures together.

- **Lack of consumable APIs.** The easiest solution to the architecture problem above is to ensure that your application provides APIs for sending in data and receiving results. That way, it can be easily integrated with your systems. However, unless your AI developer team includes full-stack engineers, they probably overlooked this requirement, or failed to provide robust APIs.
- **Security issues.** In order to function effectively, your new AI model will need to receive data inputs and provide actionable outputs. That means it needs to be tightly integrated with your backend systems. As we pointed out above, the likelihood is that your AI development team lacks software engineering experience. Therefore, there's also a high chance their code may contain security vulnerabilities.

How can you solve this?

MLOps is based closely on DevOps. One of the best definitions of DevOps is that it's a set of practices that allow you to deploy new features faster, more reliably, and with less risk. In the MLOps world, that means two things. Firstly, ensuring your team has the required skill set to create robust, secure, and reliable applications. Secondly, relying on reusable and predictable components so you don't end up reinventing the wheel each time you deploy a new solution.

Skills your team needs

A traditional AI team focused almost exclusively on data scientists. But nowadays you need a more balanced team. The following are the typical roles:

Data scientist. Data science is still a critical element for successful AI projects. You need to find hands-on data scientists with experience working on real-world projects and an instinctive feel for the data.

Mathematician. There is a lot of hardcore math involved in creating ML models. Bigger teams might want to employ someone with a background in math or physical sciences to help.

Data engineer. Data engineers are the ones that understand how to handle all your data sources, importing the data, converting it to the right format and cleaning it up.

ML engineer. This relatively new role is the one that actually builds and trains the ML models, taking into account the constraints that will exist when the model is deployed.

Full stack engineer. If you are lucky, your team will include more traditional software engineers who are able to use their experience to ensure the resulting solution is robust and deployable.

MLOps engineer. If you can afford it, you should definitely add a skilled MLOps engineer to the team. Otherwise, look to invest significant time and resources into upskilling your present team members.

Maintenance, reliability, and model drift

One of the most important tasks for MLOps is to ensure that your applications are as near 100% reliable as possible. Their aim is to minimize downtime and to build robust and fail safe systems. Your AI solution will face all the same issues that any other software does. Backend failures, network outages, external threats, and more. However, there's one way in which AI solutions are uniquely at risk. Over time, all ML models will start to drift. This is because of both internal and external factors. Put simply, when you trained the model it was trained on the world as it was then. So you need to constantly monitor the performance and accuracy of the model outputs.

What issues might you face?

The issues can be split into traditional DevOps problems and problems that are specific to AI solutions.

Maintenance. All software needs a certain amount of maintenance. This ranges from installing the latest security patches to ensuring the system is coping with the level of demand. AI applications are no exception so you need to follow standard DevOps practices, especially if the AI is mission-critical.

Reliability. Once, almost all software was installed locally. There were reliability issues, but they tended to be OS faults like the infamous blue screen of death. Nowadays, 90% of applications are delivered over the network, be that the Internet or the corporate intranet/VPN. Now, the sources of failure have become the network, configuration issues, issues in the data center, and external bad actors. AI applications need to be protected against all these, just like all your other critical applications.

Model drift. ML models are nearly all static. Once they are trained, they don't change. So, anything that causes its assumptions to change will cause your model to lose accuracy. This may be external factors, like changes in the economic climate, unseasonal weather, or new competitors entering the market. Or it could be internal factors, such as your processes and outcomes being changed by the model itself. It's critical to put in place systems that can check for model drift like this.

How can you solve this?

MLOps can solve the first two issues in just the same way as DevOps. For instance, they might give your AI dev team an error budget to work with. That means the team would be responsible for ensuring they only trigger a certain amount of system down time. They will use dashboards to monitor the performance and load on the backend and will be able to take steps to control it if there's an issue. For AI, this will include the need to monitor the health of all the data sources, both internal and external.

Managing model drift is a bit harder though. This requires complex monitoring driven by its own intelligence. Monitoring can take two forms: quantitative or qualitative. For the former, you need to regularly revalidate the model as happened during training. You probably want to look to see if the model performance metrics are getting worse. You can also check whether the outputs are differing significantly from what they were. For qualitative monitoring you can check the apparent accuracy of the model. It should soon be obvious if it is giving bad outputs. For instance, if it keeps getting predictions wrong. At this stage, you probably need to retrain the model with the latest data.

When is the right time to address AI debt?

We are often asked this question and the obvious answer is "solve it now!" But actually the real answer is more nuanced. You need to address your AI debt before it becomes too damaging to you and your business. There are a number of internal and external factors that will determine the correct timing.

	Internal factors	External factors
Available budget	Do you have any headroom in the current budget to address AI debt?	Can you raise more investment to solve AU debt?
Financial pressure	Are you seeing a reduction in your profitability?	Are competitors stealing your leads?
Efficiency	Do your teams operate inefficiently?	Do competitors seem to work smarter?
Opportunity costs	Are you losing out on deals that you expected to win?	Are competitors growing faster than you?

All of these factors and more will influence your decision. One way to help is to try and perform an ROI calculation for solving AI debt versus living with it.

The Sonasoft AI ROI equation

$$\text{AI ROI} = \sum(R_e + R_n) - \sum(C_b + C_d + C_r)$$

where:

C_b = *costs of building AI applications*

C_d = *costs of deploying an AI application*

C_r = *costs of running an AI backend*

R_e = *revenue from efficiency savings*

R_n = *revenue from new opportunities*

Should I solve AI debt by myself?

Businesses often want to solve their AI debt by themselves. But this is usually an expensive option and leads to high failure rates.

How much does it cost to develop AI?

As a rough guide, we usually put the cost of developing and running a successful AI solution at ~$2.25M. This is shown in the following table.

	Item	Total annual cost
AI developer team	Senior data scientist	**$260k**
	Mathematician x 2	**$220k**
	Senior ML engineer	**$210k**
	Junior data engineer x2	**$200k**
MLOps team	Full stack engineer x2	**$270k**
	Sr backend engineer	**$155k**
	Systems admin x2	**$210k**
	MLOps engineer	**$200k**
Backend costs	Annual cloud/platform fees	**$350k**
	Maintenance costs	**$150k**
	TOTAL COST PER YEAR	**$2.25M**

On top of these direct costs you need to account for the additional management overhead of building and maintaining a new team in-house.

What are the typical pitfalls?

Companies that set out to solve AI debt without support typically trip up in one of three ways. All of these are avoidable if you manage things really well. But that requires the entire company to buy into the process.

1. **Failures in data handling.** There are a number of issues relating to data handling that can trip your team up. For a start, there's the issue of importing all your historical data into one shared data platform. Then there's the issue of cleaning up and engineering that data without adding bias. Finally, after deployment you can hit issues when you connect the new AI solution to the live data feeds as these typically don't behave the same way as the test feeds used in development.
2. **Problems with skeptics.** Often, there is skepticism around AI solutions because so many fail. But ironically, a lot of those failures happen because of that skepticism. During the development process, the AI team will need significant support from the lines of business that will use their solution. This includes providing subject matter experts, engaging in constructive feedback, and embracing the changes that may be needed.

3 **Issues with unrealistic expectations.** The third problem is that people have been taught to view AI as almost magical. There is a broad lack of understanding about exactly what AI is or isn't capable of. This isn't helped by over-hyped news reports that imply AI is already all-powerful. Systems like GPT-3 and DeepMind can exhibit astonishing abilities and regularly hit the headlines. But these are very far from the mainstream. Only the largest multinationals can afford to use such advanced deep learning systems. For normal companies, narrow artificial intelligence must typically suffice. This can often deliver gains of 20% or more. But it can't achieve the impossible.

How can Sonasoft help you?

Sonasoft has one simple aim: to make AI as close to zero-effort as possible. Key to this is helping you solve your AI debt. We achieve this by our end-to-end approach to AI supported by SAIBRE, our AI ecosystem.

End-to-end AI solutions

Our core offer is a unique end-to-end AI approach that solves your AI debt in one fell swoop. The idea is to create a reliable AI solution that can be easily productized and maintained. This is described below.

Zero-effort data exploration

Data is king when it comes to AI. We start every end-to-end project with a data discovery and feasibility study. We help you to assemble all your data and import it into SAIBRE's data platform. Our data scientists are then able to explore the data and work with you to identify potential AI use cases. This process is rigorous and ensures we only develop AI models if there is robust data to support them.

Zero-effort model creation

Once we identify a strong AI use case, we can start to develop an AI model. Usually, we rely on supervised learning models, but we also use a whole range of other approaches. SAIBRE makes it very easy to create complex multi-stage models that include additional data processing steps. It also enables us to perform model competition to establish which model will deliver the most reliable results. It's worth noting that you often need to compromise between model accuracy, model performance, and overall deployability.

Zero-effort deployment

Often deployment is where AI initiatives fail. The team that developed the model is often blissfully unaware of the hard realities of deploying software into production. But with SAIBRE this becomes extremely simple. We offer three deployment models. You can run the model on our own infrastructure, accessing it via our APIs. Alternatively, you can take a containerized version that bundles the model and SAIBRE runtime. This can run on your own server or even on some edge devices.

Zero-effort maintenance

The final element is our smart monitoring system. This is a complete AI-powered monitoring and maintenance solution that solves all the issues around delivering reliable AI. This includes monitoring the system health, the condition of the data sources and the state of the actual model. AI models suffer from drift over time. There are various forms of model drift. But essentially they all result in the model becoming less and less accurate as time progresses. If the model becomes too poor, smart monitoring triggers automatic retraining to keep it reliable.

Other ways we help

We always try to be flexible in our approach. We understand that not every company needs a full end-to-end solution. It may be that you already have a data science team and an AI platform. Or you might just be looking to future proof your business by collecting your data in a proper data platform.

White label AI products

One of the key elements of SAIBRE is its ability to let you deploy AI solutions quickly and easily. This makes it ideal for creating white label products that you can then remarket to your customers. This is one of the best ways to monetize your valuable data. This is quite a common use case for our clients. For instance, you might have a significant collection of data on insurance cases over the last 3 decades. We can help you transform this into a model that can be sold to insurance brokers to enable them to accurately price insurance risks.

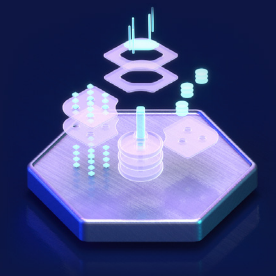
Individual SAIBRE modules

SAIBRE is designed in a completely modular fashion. That means you can use individual modules such as the data platform or smart monitoring. This can be great if you don’t need us to develop an end-to-end solution for you. For instance, some of our clients have worked with universities to develop cutting edge machine learning models. We help them to productize their solution by providing stand-alone elements of SAIBRE.

SAIBRE

Sonasoft AI Build and Runtime Ecosystem

SAIBRE is our AI ecosystem. It provides all the tools we need to deliver robust end-to-end AI solutions for our clients. SAIBRE consists of 4 elements as shown below.



Data platform

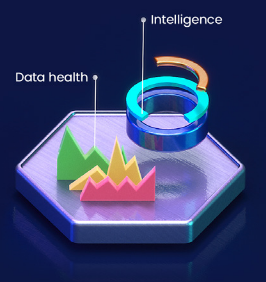
SAIBRE helps streamline the data exploration process. It allows you to import your data, or to link data from an API or 3rd party source. It provides tools for data exploration and engineering. These let you create reusable datasets for machine learning.

AI platform

This allows you to build ML models in a modular fashion. Models can be added from a growing library of templates, or via custom code. The model can then be trained and validated, or compared with other models to select the best.

AI runtime

The SAIBRE runtime engine is lightweight and efficient. Models can be deployed with just a few clicks. The resulting AI application can then be run in Sonasoft's infrastructure, or you can receive an installable container to run on your own server.



Smart monitoring

Smart monitoring uses AI to identify when your models are starting to become less accurate. If needed, it triggers model retraining. It also includes tools to monitor the quality and reliability of your data feeds.

Conclusions

AI debt has hampered many companies as they try to deliver AI transformation. Often, people fail to identify the full extent of the problem and try to solve it piecemeal. This is why so many companies out there offer you AI platforms that are designed to help relieve technical AI debt. However, you will still be stuck with data and human AI debt. That's where we come in. As you have seen, our SAIBRE ecosystem is built to directly address technical and data debt. And if you work with our highly skilled team, we will solve your human AI debt too. All this is packaged in a simple 3-step end-to-end process for delivering reliable AI solutions fast.

Sonasoft
sonasoft.com

+1 (408) 708-4000
info@sonasoft.com
1735 N. First Street, Suite 110
San Jose, California 95112 U.S.A.