AI Transparency Notices for SaaS Platforms

Nick Leghorn
Co-Founder, Secure Start Partners

Secure Start Partners

SecureStartPartners.com
info@securestartpartners.com

Introduction

Here's a scenario that I've seen pay out with increasing frequency recently.

A prospective customer is working through the process of buying an AI powered software solution from a vendor. All of the features sound great, the pricing is acceptable, and the stakeholders are all happy. But then during the security assessment portion of their purchase, the customer's security team starts asking questions about the safety and security of the software's use of Artificial Intelligence to provide the service.

The process bogs down. Sometimes the customer has a pre-conceived notion of how the AI is implemented that doesn't match reality and needs to be adjusted. Other times it becomes a lengthy sequence of questions asked back-and-forth between vendor and customer as the customer's team goes on a fishing expedition for details about the AI components of the service.

In any case, this isn't good.

The goal for any SaaS provider is to make the procurement process as quick and painless as possible: answer any questions in advance, provide the paperwork up front, and handle objections before they have a chance to derail the project. **Sales close quicker and more consistently when there are no questions left to chance**.

The good news is that we've seen this happen before. And we can use some of those same tools to try to avoid bogging down during vendor assessments and improve our sales velocity.

Hacking the Vendor Assessment Process

We live in a moment where there's significant confusion, uncertainty, and mistrust among some customers when it comes to the use of AI provided by third parties. Companies want to embrace these new technologies, but are often more willing to spend the money and resources to develop these technologies in-house rather than purchasing an external solution solely because of the potential impact to the confidentiality of their data. Providing this data to a third party could lead to privacy issues in the event of a data leak, loss of competitive advantage if the data is used to train a generally available model, or other concerns about the fairness of the model in how it would treat people and make decisions.

We've seen a similar evolution in the cybersecurity space, where companies have traditionally been wary of third parties and their security practices when handing them sensitive data. That's what led to the development of Vendor Security Assessments as a concept, a formal process for reviewing the security of a vendor before proceeding to purchase their services. These assessments started as voluntary best practices and have since been formalized as a requirement for some regulated industries.

The way SaaS platforms “hacked” their way into this process was initially through security FAQ documents and whitepapers, a practice that eventually transformed into the SOC 2 report – a standardized, consistent document prepared by a trusted third party auditor that ensured the company met the baseline security requirements that anyone would expect. Having a SOC 2 available bypassed much of the vendor security assessment process since an auditor already did the work on their behalf, closing sales quicker and preventing objections from being unnecessarily raised.

Nothing like that currently exists specifically for AI tools and systems. At some point in the future we may see the SOC 2 controls incorporate AI specific requirements, or some other audit may act as a “Get Out Of Jail Free” card for AI companies to skip this painful step in the sales process, but right now there is no trusted source for that information.

Instead of waiting for something to emerge, many companies are turning to the concept of an **AI Transparency Notice** to try and fill that role.

What Is an AI Transparency Notice

The concept seems to have originated as a compliance obligation. Proposed and upcoming regulations like the **EU AI Act** place obligations on companies to ensure that they are transparent with their use of AI and how they handle customer data, so providing some kind of a document to that effect was inevitable. But smart companies are turning this obligation into an opportunity to proactively address customer questions about AI and promote customer trust.

The core components of an AI Transparency Notice are simple, and the goal is to keep the document short and concise. We’ll walk through each of the sections of the AI Transparency Notice momentarily, but it’s important to note that when you are drafting this document there really are three audiences that should be kept in mind:

- **End users**, specifically those whose data will be included and processed as part of this system, and who want to understand how their data is used, how the decision making processes are designed, and where to report any issues.

- **Business customers** who want to make sure that the company is using their data in a safe and appropriate manner.

- **Regulators** who want to make sure that companies are being transparent about their use of AI.

Writing an AI Transparency Notice

There is currently no standard format for an AI Transparency Notice, but there are some emerging best practices that are designed to achieve all of the goals we set out for all of the potential audiences of these documents. Things to keep in mind as we dig deeper into this document:

- **Write one document per AI process**. This will keep the scope of each document clean and well organized.

- **Keep the discussion high-level**. Don't go into too much detail, especially if it might reveal any trade secrets.

Purpose and Objective of the AI System

Every system has a goal. There's a reason we created it, and specific objectives it was intended to achieve. Companies like to understand not only the outcomes but the intentions of what the AI system is designed to accomplish, and that's what this section is all about.

It's also a good place to provide a quick overview of the system in general and how it is implemented, like an introduction section to the document. You can optionally add an introduction, but in our experience it feels duplicative of this purpose and objective section and seems to simply waste time and space on the page.

Data Sources and Storage

One of the biggest concerns that users of AI systems have is where data is stored, the security of that data, and how it is segregated from other customers. That's why the biggest driver to paid AI services is the ability to manage data storage and use, attempting to reduce the possibility that sensitive company data may be leaked or used as training data.

Things to specifically call out in this section include:

- **Model Training Data** – What was the original dataset used to train the model and where was that data obtained. This is as much a question of quality as it is determining whether the customer's data is used to train the model. And if customer data is used to train the model, it might be good to highlight any anonymization steps that are taken to scrub those datasets.

- **Storage Location for Data** – Is customer data stored separately? Is each customers' data segmented from each other? Is the data encrypted? And where geographically is that data located, for GDPR concerns?

- **Data Sources for Model Use** – When the model is being used, what data is presented to it and where does it obtain that data? For example, is the system connected to your Google Drive and given access to all data within that space? Or do you need to upload specific data to use the model?

Model Training Process

AI models aren't just a product of their datasets but also the training process used during their creation. In this section the goal is to highlight the choices that were made for how to perform that training so that customers and users can understand a little better what is going on underneath the hood.

If your model is trained on a per-customer basis or has other techniques in use to limit the use of customer data or tailor each model to each customer this is the perfect place to describe those aspects of your system.
Other things to include are:

- **Process –** Generally how does the training process happen? How is the model influenced during training to achieve the desired goals?

- **Frequency –** Does the model get updated often, or is training performed on a regular basis? What would kick off a training session if it is infrequent?

- **Supervision** – Is the training a strict session performed by company employees with tightly controlled adjustments? Or does the model learn organically as it interacts with people and documents? What feedback mechanisms are available as the model is trained?

- **Customer Customization** – Can the model be trained specifically for one customer, or is it a broadly used model that cannot be customized?

Implementation and Use of AI System

Some AI systems sit and wait patiently for a prompt, like ChatGPT. Others are integrated into our systems and appear automatically, like the old Clippy function in Microsoft Word. Others sit in the background and perform tasks without anyone ever knowing. This section is intended to help customers and users of the AI system understand how it functions and is useful.

Some things to highlight in this section include:

- **In-House or Third Party? –** Do you train your own AI model, or do you use a third party like OpenAI or Gemini? If you use a third party this is where you should talk about protections you have in place, legal agreements, and technical controls that prevent data

from being used by those third parties unintentionally.

- **Any necessary API or other integrations** – Does it hook into any other systems? Will it need access to any services or datasets?

- **Access control restrictions** – How is access to this AI tool gated? Does everyone have access to it, or can that be restricted?

- **Input protection and sanitization** – If any scrubbing for PII, sensitive data, or prompt injection attacks happens, this is a good place to discuss those protections.

- **Decision Making Abilities** – One of the primary concerns with AI systems is that they might take actions or make decisions that are incorrect or might adversely impact some populations. Take some time to discuss any decisions that are made automatically by the system or to what extent humans are involved in the decision making process.

- **Outputs and Connected Systems** – Once the results are generated, where do they go? Talk about output sanitization efforts, connected systems, and storage for data once it has been processed.

Identified Risks and Mitigations

The most straightforward way to use this section is simply to talk about the risks you've identified in your platform and how you've addressed them. Transparency in this area helps encourage trust in the platform, and by displaying the thoughtful risk mitigation strategies you've implemented it can indicate that you'd likely take the same approach to anything not explicitly listed.

But there's another reason to expand on this section as well.

The SOC 2 audit has a concept called a "Complimentary User Entity Control" which is a fancy way of saying "stuff customers need to do to protect themselves while using this service." Cloud services do the same thing with their "shared responsibility model" that they publish.

The same concept applies here. No system is perfect, and in some cases we expect customers to put in place basic security controls on their end to prevent security issues. For example, if you allow people to create a unique username and password to log into your service, you should probably remind people that the confidentiality and security of that credential is their responsibility.

Not only will it be a good reminder of people to take their own security seriously, but it can also be great evidence to limit your legal liability should something happen down the road.

Reporting Errors and Other Issues

This section should be short and to the point: if something goes wrong with the system, such as an incorrect outcome or an output looks wrong, where can the end user of the system report that? It should be an email address that you'd expect to get similar notices, such as security@company.com or support@company.com .

Conclusion

We're in a Wild West moment for AI systems, and until a consistent approach for doing vendor management and risk management for these systems appears similar to how the SOC 2 audit exploded in popularity we're going to keep seeing sales deals get bogged down during the security assessment portion of the deal.

In the meantime, our recommended strategy for improving sales and maintaining compliance with emerging regulations is to be transparent about the use of AI, how it is trained, data sources that it uses, and how the resulting data is used by the business. The best way to do that is through an AI Transparency Notice, which we have described here and will provide an example below for you to follow and reproduce.

AI Transparency Notice for DeliveryDetective.com

Purpose and Objective of the AI System

The DeliveryDetective.com platform is designed to provide a better interface and system for tracking packages sent through multiple shipping agencies. The AI system within our platform uses past delivery history and current weather forecasts to predict the likelihood that a package will be delayed beyond the current delivery window and provide that as visual feedback to the end user of our platform.

Data Sources and Storage

All data obtained and stored within our platform is located within the United States, specifically within the DreamHost cloud hosting platform located in California. Data is encrypted at rest and in transit, and strong protections are in place to ensure that only appropriately authenticated and authorized individuals and systems are permitted to access any non-public data.

Our data includes:

- **Package Tracking Data** obtained from shipping companies. This generally includes the rough geographic locations of packages, including source and destination. This includes historical data as well as real-time data for packages being tracked in our system.
- **Historical and Real Time Weather Data** obtained from open source and government sources.
- **Real Time Traffic Data** provided by Google Maps.
- **Geographic Data** including distance between locations.

Model Training Process

Our AI model is continuously trained on the data we obtain. Our systems are designed to accept inputs related to package location, traffic, current weather, and expected delivery window, and generate a confidence score related to whether the package is likely to arrive within the stated window. Our systems automatically provide feedback in the form of whether the confidence window was accurate, which is used to improve the reliability of our models.

Implementation and Use of AI System

Our AI system is automatically triggered whenever a new package is added to the system for tracking, and will automatically generate a confidence score related to the delivery of that package.

Inputs are provided from only trusted sources, including the original source of truth for tracking data for packages, weather information, and other similar sources. All data is authenticated, sanitized, and validated prior to being used within our models. The only output available from our model is the confidence rating for the delivery which is displayed within our dashboard.

Our models are not generative, do not accept text prompts from users of our systems, and do not output any data that could include sensitive information. No automated actions are taken based on the results of our model.

Identified Risks and Mitigations

As with all technology, there are some risks associated with the use of advanced AI systems such as this. We have identified the following risks related to our use and management of this AI system and have divided those risks into areas where the company has taken steps to reduce those risks and areas where deployers of the AI system should implement some complimentary controls to appropriately manage those risks.

Provider Managed Risks

- All inputs and outputs are sanitized and validated before being accepted, used, or displayed back to the user of our platform.

- Provider manages the API keys and credentials for the data sources used within the model and ensures the confidentiality of the information provided by end users. Data is only stored within the platform, and agreements exist to prevent data storage or unauthorized use of data by third parties.

Deployer Managed Risks

- Deployers and users of our system are reminded that they are responsible for maintaining the confidentiality of their own credentials used to access the system.

Reporting Errors and Other Issues

For any identified errors, issues, inaccuracies, or other requests related to the safety and security of this AI system, please contact: info@deliverydetective.com