



Responsible AI Statement

e: info@redactxpert.ai

w: <https://www.redactxpert.ai>

Powered by Simpson Associates



Contents

- 1. Responsible AI 3
 - 1.1 Human-in-the-Loop..... 3
 - 1.2 How the AI services work..... 3
 - 1.2.1 Pattern matching 4
 - 1.3 Data security and processing 4
 - 1.4 Getting the best results from AI detection 4
 - 1.5 Global identification..... 5
 - 1.6 Restricted use cases 5
 - 1.7 Model training..... 5

1. Responsible AI

This Responsible AI Statement sets out how RedactXpert uses artificial intelligence to assist with the identification and redaction of personally identifiable information (PII) in documents. It describes the principles and safeguards that underpin the product to provide transparency, fairness, and explainability, including the requirement for human oversight of all redaction decisions, how the AI services operate and handle customer data, guidance on getting the best results from the AI detection process, and the limitations and restrictions that apply to the tool's use.

RedactXpert is committed to the responsible and ethical use of AI. The AI within this product is a tool to support human decision-making — not to replace it.

1.1 Human-in-the-Loop

RedactXpert is designed with human oversight at its core. The AI services within the application perform automatic identification of potential PII, but at no point does the system automatically apply redactions without explicit user confirmation. Every item flagged by the AI model must be reviewed and approved by a human operator before any redaction is applied to a document. This ensures that all redaction decisions remain under human control and that no sensitive information is removed without a deliberate action from a user.

RedactXpert provides a selection view mode which draws an outline around the selected redactions, so users can easily distinguish which areas of the document will be redacted. This can be used to allow further human review where necessary.

1.2 How the AI services work

RedactXpert processes documents with AI in two stages. First, OCR (Optical Character Recognition) is performed on uploaded documents to extract text and its positional information. This extracted text is then passed to an AI language model to detect potential PII within the content.

The auto-detection of PII uses context to determine whether a particular value should be highlighted as PII or not. It is much more likely for PII to be detected if it has appropriate context around it. For example, a phrase such as "Mark is 35 years old" provides context that allows the AI model to more confidently determine that "Mark" is a name, whereas "Mark" appearing in isolation would be less likely to be identified (there is no fixed list of names used for matching). The findings from the AI process are recommendations only and must be reviewed by a human before they are included. No decisions are directly made by the AI models without human input.

RedactXpert uses Microsoft Azure AI Foundry tools. For further information on Microsoft's responsible AI approach, please refer to Microsoft's responsible AI overview: <https://learn.microsoft.com/en-us/azure/foundry/responsible-use-of-ai-overview>

1.2.1 Pattern matching

In addition to the AI language model, RedactXpert also uses pattern matching to supplement PII detection. Pattern matching works by identifying values that conform to known formats or structures - for example, UK Driving Licence Numbers - and flagging these as potential PII regardless of the surrounding context. This allows the system to detect certain categories of sensitive information that may not be reliably identified by the AI model alone, providing an additional layer of coverage. As with all findings from the AI process, any matches identified through pattern matching are presented as recommendations only and must be reviewed and confirmed by a human before any redaction is applied.

1.3 Data security and processing

Sensitive data uploaded by customers is transmitted to Microsoft Azure AI services for processing. Security is implemented at every stage of this process. Information on how the Azure AI Language model handles data, including privacy and logging considerations, can be found in Microsoft's documentation:

<https://learn.microsoft.com/en-us/azure/ai-foundry/responsible-ai/language-service/data-privacy?view=foundation-classic>

In particular, the LoggingOptOut parameter controls whether temporary logs are stored on Microsoft's side during processing. Although the default behaviour for the PII model disables such logging, RedactXpert explicitly sets this parameter to true as part of its processing pipeline, opting out of storing any logs with Microsoft.

1.4 Getting the best results from AI detection

To maximise the accuracy and reliability of PII detection, the following guidance is recommended:

- **Document clarity:** Ensure documents are clear and legible before uploading. Poor scan quality, low-resolution documents, or small font sizes may reduce OCR accuracy, which in turn can affect PII detection.
- **Context-driven:** As described in "How the AI services work", the AI model uses context to determine potential PII. Best results will be obtained when there is context surrounding the PII.
- **Tables and context:** Tables are used to infer context through column headers. Using descriptive column headers will enable the AI model to better understand the nature of the data and produce more accurate results. Documents with clearly defined table borders are also likely to yield better results as the AI model will be able to determine more accurately where rows and columns start and end.
- **Excel files:** Excel files are converted to PDF using the print preview layout equivalent to opening the file and choosing Print to PDF. You can improve results by adjusting the print layout beforehand to ensure the document is readable and that tables appear on relevant pages rather than being split across page breaks. In particular, ensure that no columns are cut off; if a cell's content is obscured by an overlapping adjacent cell, that text may not be fully read by the OCR process.

1.5 Global identification

The AI language model used by RedactXpert is a global model, trained on a broad and diverse range of data rather than being limited to any single region or jurisdiction. This means the model is capable of identifying PII across a wide range of categories and formats as they appear in different countries and contexts, not solely those common to UK documents. As a result, the model is better equipped to handle documents that contain international data, names, identifiers, and conventions, reducing the risk of certain categories of PII being systematically overlooked due to regional bias in the underlying training data.

Users should nonetheless remain aware that, as with any AI model, detection accuracy may vary across different PII categories, languages, and document types. Human review of all findings remains essential to ensure completeness and accuracy.

1.6 Restricted use cases

RedactXpert and its underlying AI services are intended to assist human reviewers in identifying PII within documents. The tool is explicitly not designed or intended to be used as an auto-redaction solution. All AI-generated findings are recommendations that require human review and confirmation before any action is taken.

Additionally, documents containing Type 3 fonts are not supported. Type 3 is a legacy font format that can exhibit unpredictable behaviour when redactions are applied, potentially resulting in a larger area of the document being removed than intended. A warning will be displayed to users in RedactXpert when they enter a document that has been identified as containing Type 3 fonts.

1.7 Model training

No data from documents uploaded by customers is used to train AI models. Customer data processed through RedactXpert is used solely for the purpose of PII detection and is not retained (outside of the configured retention period), repurposed, or shared for model training or improvement by RedactXpert or its underlying AI service providers.

While customer data is never used for model training, RedactXpert works closely with Microsoft to support the ongoing improvement and development of the underlying AI models. This includes providing feedback on new categories and formats of PII, emerging data patterns, and areas where detection accuracy can be enhanced. This collaborative approach helps to ensure that the model continues to evolve in line with real-world use cases, without any compromise to customer data privacy.