# AI Transparency Statement

## Can I Phish Pty Ltd

## 1. Purpose & Scope

This AI Transparency Statement explains how Can I Phish Pty Ltd ("CanIPhish", "we", "our") uses artificial intelligence (AI) within the CanIPhish platform. It is intended to help our customers, their employees, and any relevant regulators understand the nature of the AI systems we deploy, how those systems are governed, and the safeguards we have put in place.

This statement applies to all AI features within the CanIPhish platform as at the effective date above, and is updated whenever material changes are made to our AI systems or governance practices.

This document has been prepared with reference to the EU Artificial Intelligence Act (Regulation (EU) 2024/1689) and represents CanIPhish's good-faith effort to operate as a responsible AI deployer.

## 2. Company Overview

**Company:** Can I Phish Pty Ltd (ABN 24 647 802 266)
**Headquarters:** Australia
**Product:** Phishing Simulation & Security Awareness Training Platform
**Website:** https://caniphish.com
**Privacy Contact:** privacy@caniphish.com

CanIPhish provides organisations with tools to simulate real-world phishing attacks in a safe, controlled environment and to train employees to recognise and respond to cybersecurity threats. Our platform is used by managed service providers (MSPs), enterprise customers, and organisations of all sizes across more than 65 countries.

## 3. AI Systems Used in the Platform

CanIPhish incorporates AI across seven distinct features. Each is described below, including its purpose, the AI services involved, data processed, and applicable safeguards.

## 3.1 Business Impact Rule Generation

The Business Impact feature fine-tunes the Human Risk metric by applying AI-generated rules to assess the potential business impact of an employee falling for a phishing attack. Rules are generated using AI based on job title data in aggregate, no individual employee is profiled or scored by the AI itself.

- **Purpose:** Generate business impact rules that contribute to an organisation's Human Risk Metric scoring.
- **Subprocessor:** Microsoft Azure OpenAI.
- **Data processed:** Job titles only, processed in aggregate. No individual employee personal data is sent to the AI service for this feature.
- **Data location:** Configurable. Azure OpenAI processing occurs in the same region as the customer's chosen data storage location (with the exception of Singapore tenants, which process via Japan, and UAE tenants, which process via Germany).
- **Human oversight:** Administrators review and apply generated rules. The AI does not make final scoring decisions autonomously.

## 3.2 AI Content Generator (Training Modules, Simulated Phishing Emails & Simulated Phishing Websites)

The AI Content Generator enables administrators to create custom security awareness training modules, simulated phishing emails, simulated phishing landing pages/websites, and AI-generated training videos. Administrators can upload organisational policies, provide custom AI prompts, and supply content parameters to tailor simulated attacks or training materials to their environment.

For phishing simulations, the AI is used solely to generate realistic but controlled phishing email content and phishing website text and structure for use within authorised campaigns. No real credentials harvested during simulations are retained or used by the AI.

- **Purpose:** Generate custom security awareness training modules, simulated phishing emails, simulated phishing websites, and professional-quality training videos tailored to the customer's policies and campaign objectives.
- **Subprocessor(s):**
    - Microsoft Azure OpenAI (text generation for training content, simulated phishing emails, and simulated phishing websites)

    - ElevenLabs (text-to-speech audio for training videos only)
    - Fal AI (video generation orchestration for training videos only)
    - Veed (underlying video AI model for training videos only)
- **Data processed:**
    - For training and phishing content generation: campaign prompts, policy text, and simulation parameters provided by administrators.
    - No employee personal data is processed by AI when generating simulated phishing emails or simulated phishing websites.
    - For video generation, script transcripts may be processed by ElevenLabs, Fal AI, and Veed. These subprocessors do not retain data beyond the processing
- **Data location:**
    - Azure OpenAI: Configurable, matching the customer's chosen data storage region
    - ElevenLabs, Fal AI, and Veed: United States
- **Human oversight:** All training content, simulated phishing emails, and simulated phishing websites are reviewed, approved, and launched by an administrator. The AI does not autonomously publish or deploy content.

## 3.3 Voice Phishing Simulator

The Voice Phishing Simulator delivers AI-powered outbound phone calls to employees as part of simulated vishing (voice phishing) campaigns. Calls are context-aware, conducted in real time, and can include deepfake voice capabilities where administrators upload a custom voice sample. All voice phishing simulations operate under a strict double opt-in consent framework.

- **Purpose:** Simulate realistic voice phishing attacks to train employees to identify and resist phone-based social engineering.
- **Subprocessor:** ElevenLabs (conversational AI voice agent and text-to-speech.
- **Data processed:** During an active call: real-time audio stream, live call transcript, call metadata (timestamps, duration, session ID), and AI prompts, which may include employee name and job title. No audio or transcripts are retained by ElevenLabs or CanIPhish after the call ends.
- **Data location:** ElevenLabs process data in the United States.
- **Consent:** Mandatory double opt-in via SMS before any employee receives a simulated vishing call.

## 3.4 Conversational Email Phishing

PhishAI is CanIPhish's conversational email phishing engine. It simulates a technique used by real-world attackers in which an initial phishing email is sent to establish contact, and the attacker then engages the victim in a back-and-forth email conversation to build trust before delivering a malicious payload. The AI reads employee responses and generates contextually appropriate counter-responses, ultimately determining whether the employee has been successfully phished.

- **Purpose:** Simulate multi-turn, AI-driven email phishing conversations to train employees to recognise sophisticated social engineering attempts.
- **Subprocessor:** Microsoft Azure OpenAI.
- **Data processed:** Employee email responses are read by the AI to generate counter-responses and assess whether the phishing payload was accepted. Employee information including email address, first name, last name, job title, company name, and residing country may be shared with Azure OpenAI during an active campaign.
- **Data location:** Configurable. Azure OpenAI processing occurs in the region matching the customer's data storage location (Singapore tenants process via Japan; UAE tenants process via Germany).
- **Disclosure:** Upon a successful phishing outcome, the employee immediately receives notification that they were part of a simulated phishing exercise, followed by targeted micro-learning.

## 3.5 AI-Powered Report Email Plugin & Threat Analysis

CanIPhish provides an email add-in that allows employees to report suspected phishing emails directly from their inbox. Reported emails are automatically sandboxed and subjected to AI-powered threat analysis to classify them as a real threat, a known phishing simulation, or a false positive. The AI can also perform embedded analysis directly within the email client interface.

- **Purpose:** Enable employees to report suspicious emails and provide AI-assisted triage and threat classification to administrators.
- **AI service:** Microsoft Azure OpenAI.
- **Data processed:** Content of employee-reported emails, including subject line, sender address, body text, URLs, and attachments. Analysis is performed within a sandboxed environment.
- **Data location:** Configurable, matching the customer's chosen storage region.

- **Human oversight:** AI analysis results are presented to administrators as informational outputs. All response and remediation decisions remain with the human administrator.

## 3.6 AI Program Manager

The AI Program Manager enables administrators to use AI to recommend, create, and manage an ongoing security awareness program, including simulated phishing and training campaigns. Through a conversational, back-and-forth interaction with a fine-tuned AI model, the feature can plan campaign cadence, vary phishing difficulty, and surface program improvement recommendations over time. Administrators control the level of autonomy granted to the feature: by default it operates in a recommend-and-approve mode, and any autonomous creation or launching of campaigns occurs only within boundaries explicitly configured by an administrator.

- **Purpose:** Recommend, create, and manage an ongoing security awareness program—including simulated phishing and training campaigns—tailored to the organisation's risk profile and program objectives.
- **Subprocessor:** Microsoft Azure OpenAI.
- **Data processed:** Tenant-level program configuration, campaign history, and aggregate program performance and adoption metrics. Employee-level performance signals (such as simulation outcomes) may inform campaign sequencing within the customer's tenant. No customer data is used to train any AI model.
- **Data location:** Configurable. Azure OpenAI processing occurs in the region matching the customer's chosen data storage location (Singapore tenants process via Japan; UAE tenants process via Germany).
- **Human oversight:** Administrators define the AI Program Manager's level of autonomy. By default, recommendations require administrator approval before any campaign is created or launched. All actions taken by the feature are recorded in the tenant audit log, and administrators can adjust or disable autonomous operation at any time.

## 3.7 AI Chat Widget

The AI Chat Widget is an in-platform assistant, accessible from the bottom-left of the screen when using the CanIPhish platform, that helps users with general questions about the platform. To provide relevant, context-aware answers, the assistant has visibility at the tenant level into

the tenant audit log and into broad, high-level feature adoption across the platform. No tenant-level metadata is stored or used to train AI models.

- **Purpose:** Provide users with conversational, in-platform assistance and answer general questions about platform usage and feature adoption.
- **Subprocessor:** Microsoft Azure OpenAI.
- **Data processed:** User questions, together with tenant-level audit log entries and high-level feature adoption data, used to generate context-aware responses. No tenant-level metadata is stored or used to train AI models.
- **Data location:** Configurable. Azure OpenAI processing occurs in the region matching the customer's chosen data storage location (Singapore tenants process via Japan; UAE tenants process via Germany).
- **Human oversight:** The AI Chat Widget provides informational assistance only. It does not take actions or make configuration changes on the platform, and users remain responsible for any actions they choose to take based on its responses.

## 4. EU AI Act Risk Classification

The EU Artificial Intelligence Act (Regulation (EU) 2024/1689) applies a risk-based framework to AI systems. CanIPhish has assessed each of its AI features against this framework.

| AI Feature | Risk Tier | Rationale |
|---|---|---|
| Business Impact Rule Generation | Minimal Risk | Processes only job titles in aggregate. Does not profile individuals or make consequential decisions about specific employees. Human administrator reviews all generated rules. |
| AI Content Generator | Minimal Risk | Processes policy documents and scripts provided by administrators, not employee personal data. All outputs reviewed by an administrator before publication. No consequential decisions about individuals. |
| Voice Phishing Simulator | Limited Risk | Operates under mandatory double opt-in consent. No audio or transcripts retained post-call. No biometric identification or emotional recognition performed. Transparency provided to employees post-simulation. |
| Conversational Email Phishing | Limited Risk | Entirely opt-in. Used exclusively for consensual security awareness training. Does not make consequential decisions about employees. Transparency is provided immediately upon successful simulation. |

| AI Feature | Risk Tier | Rationale |
|---|---|---|
| AI-Powered Report Email Plugin & Threat Analysis | Minimal Risk | Decision-support tool only. Final threat response decisions remain with human administrators. Does not affect employee rights or employment outcomes. |
| AI Program Manager | Minimal Risk | Operates exclusively within the consensual security awareness training context. Autonomy is administrator-configured and bounded, with human approval required by default and all actions recorded in the tenant audit log. Makes no consequential decisions affecting employees' rights or employment outcomes. |
| AI Chat Widget | Minimal Risk | Informational assistant only. Answers questions using the tenant audit log and aggregate adoption data, and takes no actions on the platform. No tenant metadata is stored or used for model training, and no consequential decisions are made about individuals. |

None of CanIPhish's AI features fall within the "high-risk" categories defined under Annex III of the EU AI Act (which covers areas such as biometric identification, employment screening, law enforcement, and critical infrastructure). Accordingly, CanIPhish is not subject to the conformity assessments, technical documentation obligations, or registration requirements applicable to high-risk AI providers.

CanIPhish operates as an AI "deployer" under the Act. We do not develop foundational or general-purpose AI (GPAI) models from scratch; we integrate and fine-tune existing AI capabilities within a specialised application context.

# 5. AI Governance & Oversight

CanIPhish is committed to responsible AI deployment. The following governance mechanisms are in place.

## Human Oversight

Every AI feature within CanIPhish operates within parameters defined and controlled by a human administrator. No AI system on the platform takes autonomous action without an administrator having first configured the campaign, selected the target employees, and initiated the simulation. Where the AI Program Manager is configured to operate autonomously, it does

so only within boundaries explicitly defined by an administrator, and all actions it takes are recorded in the tenant audit log and can be reviewed, adjusted, or disabled at any time.

## Consent & Opt-In Controls

Voice phishing simulations require explicit double opt-in from participating employees, verified via SMS. The conversational email phishing engine is entirely opt-in at the campaign level. No data is shared with any AI subprocessor unless the relevant feature has been explicitly enabled by an administrator.

## Transparency to End Users

Employees are always informed after a successful simulation that they were the subject of a security awareness exercise. No simulation is designed to cause lasting deception or psychological harm. Educational micro-learning content is provided immediately following any simulated compromise.

## Data Minimisation & Privacy by Design

CanIPhish's AI systems are designed to process the minimum data necessary. No employee audio or call transcripts are retained after a voice phishing call ends. AI subprocessors that handle video or content generation (Fal AI, Veed) receive only script transcripts, not employee personal data, and do not retain data beyond the processing request. No customer data is used to train any AI model without explicit consent.

## AI Literacy

CanIPhish staff responsible for developing, maintaining, and operating AI features receive ongoing training on the capabilities and limitations of the AI systems in use, consistent with Article 4 of the EU AI Act.

## Subprocessor Governance

All AI subprocessors are subject to contractual data processing agreements that restrict their use of CanIPhish customer data, prohibit model training on that data, and require compliance with applicable privacy laws. Subprocessors are regularly reviewed and documented on the CanIPhish Security & Compliance page at caniphish.com/security.

### Incident Response

CanIPhish maintains an internal process for identifying, escalating, and responding to unintended AI behaviour or outputs. Customers who observe unexpected AI behaviour are encouraged to report it to privacy@caniphish.com.

## 6. Subprocessor Summary

The following table summarises the AI subprocessors used by CanIPhish and the data they process. Full subprocessor details, including contractual safeguards, are published at caniphish.com/security.

| Subprocessor | Used For | Data Processed | Data Location |
|---|---|---|---|
| **Microsoft Azure OpenAI** | Business Impact Rule Generation; Conversational Email Phishing; AI Threat Analysis; AI Content Generator (training content, simulated phishing emails, simulated phishing websites); AI Program Manager; AI Chat Widget | Job titles (aggregate); employee email responses & personal data (during active phishing campaigns); reported email content; tenant program configuration, audit log entries, and aggregate adoption metrics (AI Program Manager and AI Chat Widget). No tenant metadata stored or used for model training | Configurable |
| **ElevenLabs** | Voice Phishing Simulator (AI voice agent); AI Content Generator (text-to-speech audio) | Real-time call audio stream & transcript (not retained post-call); call metadata; AI prompts (may include name & job title, deleted post-call); video script transcripts | United States |
| **Fal AI** | AI Content Generator (video generation orchestration) | Video script transcripts only. No employee personal data. Not retained beyond processing request. | United States |
| **Veed** | AI Content Generator (underlying video AI model) | Video script transcripts only. No employee personal data. Not retained beyond processing request. | United States |

## 7. Prohibited & Out-of-Scope Uses

CanIPhish's AI features are designed exclusively for legitimate security awareness training. The following uses are explicitly prohibited under our Terms of Service.

- Using CanIPhish to conduct phishing simulations against individuals who have not consented and are not employees, contractors, or authorised targets of the subscribing organisation.
- Using the platform to harvest real credentials or sensitive information from employees.

- Using voice phishing simulations against individuals who have not completed the mandatory double opt-in consent process.
- Using AI-generated content (training videos, modules) for any purpose other than employee security awareness education.
- Using the platform for any purpose other than lawful security awareness training and employee education.

CanIPhish does not use, and does not permit customers to use, the platform in ways that would constitute prohibited AI practices under Article 5 of the EU AI Act, including subliminal manipulation, exploitation of vulnerabilities, or social scoring.

## 8. Customer Obligations

As deployers of the CanIPhish platform, our customers share responsibility for ensuring that simulations are conducted ethically and legally. Customers are responsible for:

- Ensuring they have the lawful authority and, where required, legal basis to conduct phishing simulations against their employees.
- Complying with applicable employment law, privacy law, and any collective bargaining agreements before launching simulations.
- Ensuring employees participating in voice phishing simulations have completed the mandatory double opt-in consent process before any call is initiated.
- Ensuring that AI-generated training content is reviewed and approved before being published to employees.
- Using the platform's built-in compliance reporting features to maintain records of simulations for audit purposes.

CanIPhish provides built-in compliance reporting that customers can use to demonstrate to auditors, regulators, or their own stakeholders that security awareness training activities have been conducted in accordance with applicable requirements.

## 9. Updates to This Statement

This AI Transparency Statement will be reviewed at least annually and updated whenever CanIPhish introduces material changes to its AI systems, data practices, or governance framework. The current version of this statement is always available at caniphish.com/ai-transparency.

Customers will be notified of material updates via email or platform notification at least 30 days prior to the changes taking effect.

## 10. Contact

For questions about this AI Transparency Statement, our AI governance practices, or to report a concern relating to AI behaviour on the platform, please contact:

**Email:** privacy@caniphish.com
**Website:** https://caniphish.com/contact-us
**Postal address:** Can I Phish Pty Ltd, Australia

*Last Updated: June 22, 2026*

*This document does not constitute legal advice. CanIPhish recommends that customers seek independent legal counsel regarding their own obligations under the EU AI Act or any other applicable regulation.*