K2's Privacy Issues & Solutions - *What You Need To Know!*

K2 Enterprises

What About Randy?

- 40+ years of technology experience, top-rated speaker for almost 40 years
- Top 25 Thought Leaders in Accounting 2011-2024
- 2004-2023 Accounting Today 100 Most Influential in Accounting for twenty years
- Inducted Accounting Hall of Fame, Feb 2011
- Monthly columns on technology in CPAPractice Advisor, weekly podcasts on technology
- Published author of six books, From Hutchinson, KS
- randy@k2e.com or randyj@nmgi.com
- 620-664-6000 x 112

Course Description

Generative Artificial Intelligence captured the imaginations of everyone when it was launched in 2022, but they also create opportunities for data leakage and inappropriate use of confidential data based on the terms of service. Unfortunately, everyone has at some point clicked "I agree" on a terms of service (TOS) or end user license agreement (EULA) document which they neither read nor understood. This session will review the common terms embedded in these documents with thousands of words and incomprehensible terminology. Attend this breakout and learn what you wish you knew about analyzing and evaluating license agreements and privacy policies.

Major Topics

- Tools for evaluating and summarizing licenses and privacy policies
- Definitions of major terms and concepts which are used in software licenses
- Review key terms associated with major services like Microsoft 365 or QuickBooks Online

Learning Objectives

- Identify at least one tool which helps users to organize and identify key sections of license agreements
- From a list, select the correct definition for key terms used in license agreements and privacy policies
- Explain how to identify key sections in license agreements, including how to "opt out" of features which may compromise your privacy

Overview

- Software licenses and privacy policies
- Artificial intelligence laws and regulations
- Key terms from NIST
- Language models
- AI Risk management

SOFTWARE LICENSES AND PRIVACY POLICIES

What Is Software Licensing?

- Intellectual property rights have always been a problem in the most advanced societies of the world (e.g., Dutch China patterns)
  - Criminal prosecution of property crimes seems to have dropped off
  - Civil recovery seems to be more popular, with fees/fines and/or asset forfeiture
- Software piracy is protected under copyright in the US and Canada, and provides severe civil and criminal penalties for those who are unauthorized to use the software in the manner which has been permitted
- With the advent of data as a marketable product, the privacy risks associated with both AI and software usage must be managed effectively or data might leak out of systems

About Software Licensing

Software licensing is like airfare pricing in many ways:

- Its objective is to make it possible for the company/seller to maximize its revenue by segmenting its customers
- There are special deals for big customers which are not available to small customers
- The terms of what is allowed for each product segment vary depending on what you bought, how much you paid, and the sales needs of the licensor

Why Software Licenses Are Critical

- To better understand the risks associated with privacy and AI, we need to **READ** and **UNDERSTAND** the major documents associated with the application
- Two of the key documents associated with your rights/license to use software or cloud-based services include:
  - End User License Agreement (EULA) / Terms of Service (ToS)
  - Privacy Policy
- There are also common contract terms that need to be understood, particularly as you are using third-party hardware, software, and AI models to process confidential information

Note About Software Licensing

- There are numerous bundles and SKU's which change frequently
- Even the "experts" struggle to explain the logic underlying the pricing strategies
- Every software company writes its own contracts and license conditions
- We will focus on three basic documents/contracts related to your contractual licensing of software:
  - End User License Agreement
  - Terms of Service
  - Privacy Policy

End User License Agreement (EULA) / Terms of Service (ToS)

- Document that states the terms which govern your use of an application
- Let software companies define the terms for the use of their software or service in a way which is favorable to their business
- Are often accepted without ever reading them at all
- May incorporate other documents, policies, and terms by reference

Word Count/Complexity of EULAs

- Intuit QuickBooks Online
  - 57,053 words
  - Flesch-Kincade grade 14.6
- QB Desktop US
  - 24,619 words
  - Flesch-Kincade grade 14.1
- Xero
  - 5,332 words
  - Flesch-Kincaid Grade 11.8

Many EULAs are long, complex, difficult to understand, and often disclose as few specifics as possible about any practices which might be a concern to end users

Key Terms In EULA/ToS Documents

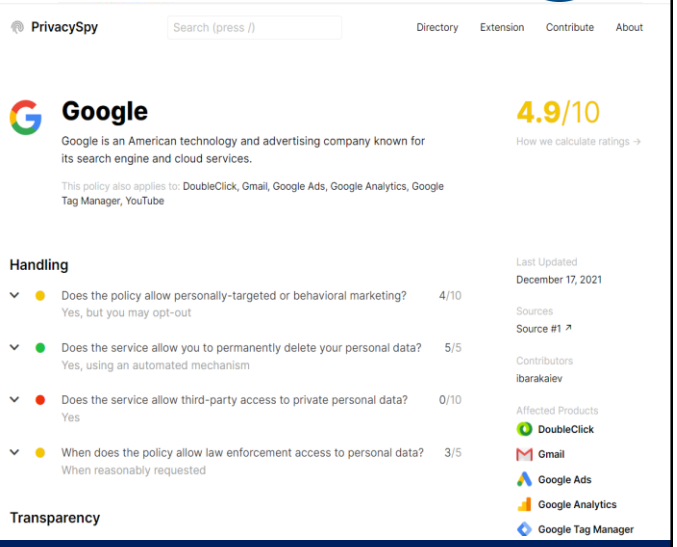
- EULAs are more commonly used for the licensing of software
- ToS are more commonly used for services hosted on websites
- Key terms/clauses include:
  - Terms and scope of license – geography, exclusivity
  - Duration of the license
  - Permitted uses and prohibited uses of the application
  - Licensing fees, subscriptions, or royalties for use
  - Disclaimers of liability and warranties
  - Limitation of liability
  - How changes are made to the EULA/ToS

Key Terms In EULA/ToS Documents

- Key terms (continued)
  - How to get help with the application
  - Dispute resolution procedures
  - Required binding arbitration
  - Jurisdiction for dispute resolutions
  - Procedures for terminating relationship
  - Consent to send communicate via e-mail, SMS, and telephone
  - Any required indemnifications (guarantees)
  - How data will be shared with others, such as sub-processors of data (e.g., bank feed aggregators)

Privacy Policy Analysis Resources

- EULAlyzer Pro – software to analyze privacy policies and terms of service
  - Pro version is only $19.95
- **ToS;DR** (Terms of service; didn't read) – tosdr.org – database of privacy ratings of privacy policies of popular sites
- **Privacy Spy** – web database of site privacy evaluations for major websites – privacyspy.org



Privacy Policy

- Document that discuses a company's privacy policy, including data like:
  - What data is gathered from you
  - How your information will be used
  - The purpose of gathering that data and how your data may be used
  - When your data will be disclosed others, including sensitive personal data and information
    - Not always specific on with whom it will be shared
  - Security policies and procedures

Word Count - Privacy Policies

- Intuit 12,221
- Xero 2,671
- ADP
  - General 13,288
  - Glossary 2,685
  - For client employees 912
- Yodlee Envestnet
  - 5,971 words
  - Flesch-Kincaid grade 14.4

Privacy Policy Extracts - Intuit

- *Intuit Privacy Policy last updated 9/6/2023 (retrieved and was current as of 3/22/2024)*
  - "To the extent we allow you to input personal information (as the term is defined under applicable law) about other individuals other than yourself, you represent and warrant that you have complied with all applicable laws and received the proper authority or consent to allow us to collect and process such information to operate our business, in accordance with our Global Privacy Statement. You further agree that, **other than with respect to information furnished to TurboTax in connection with the preparation of an individual tax return, any sharing of personal information among Intuit Group Companies is contemplated as part of the Platform**."

Intuit's Subsidiaries include:

- *Credit Karma*
  - *Credit Karma Offers, Inc.*
  - *Credit Karma Insurance Services, LLC*
- *Intuit Financing, Inc.*
- *Intuit Insurance Services, Inc.*
- *Intuit Mortgage, Inc.*
- *Intuit Payroll Holding, LLC*
- *Intuit Payroll Services, LLC*
- *Intuit Sales Tax, LLC*
- *Lacerte Software Corporation*
- *Paycycle, Inc.*
- *Payroll Solution, Inc.*
- *Tsheets.com, LLC*

Complete list is online here

Privacy Policy Extracts - Intuit

- *Intuit Privacy Policy last updated 9/6/2023 (retrieved and was current as of 3/22/2024)*
  - "The Platform may also include information about or offers for third-party services or products or allow you to connect your account to or otherwise access third-party services or products. Intuit does not warrant, and is not responsible for, such third-party services and products or claims made about them or the actions or inactions of any third party. You must review and comply with any Additional Terms. Intuit may be compensated by those third parties, which could impact whether, how and where the services and products are displayed."

Privacy Policy Extracts - Intuit

- Your Content remains yours, which means that you retain any intellectual property rights that you have in your Content. By sharing your Content on the Platform, you hereby grant Intuit a license to use your Content, as described in more detail below.

1. What's covered - This license covers your Content to the extent your Content is protected by intellectual property rights.
2. Scope - This license is:
   - Worldwide, which means it's valid anywhere in the world;
   - Non-exclusive, which means you can license your Content to others; and
   - Royalty-free, which means there are no fees for this license
3. Rights - This license allows Intuit to:
   - **Host, reproduce, distribute, communicate, sublicense and use your Content** — for example, to save your Content on our systems and make it accessible from anywhere you go;
   - **Publish or publicly display your Content, if you've made it visible to other**s; and
   - **Modify and create derivative works based on your Content**, such as reformatting or translating it

Privacy Policy Extracts - Xero

- **29. Yodlee and bank feeds:** *Your use of automated bank account feeds enabled by Yodlee from within our services is subject to separate Yodlee terms*. If your bank or credit union connects to Xero directly, you may use those feeds instead of Yodlee. Bank feeds are generally offered for free but may have associated charges that we will pass on to you. You may discontinue the use of a bank feed at any time - check out how to stop a bank feed on Xero Central. You can learn more about bank feeds on Xero Central. If you receive a bank feed and are based in the United Kingdom, Xero's additional terms for account information services apply.

Most software providers use sub processors like Yodlee or Fiserv to handle tasks like establishing and maintaining bank feeds. The use of those services may result in information being shared in accordance with additional terms and services

Yodlee's Privacy Notice is an additional 3,812 words, and the organization's Privacy Policy is an additional 2,158 words.

ADP's AI Ethics Statement

ADP's privacy statements include an AI ethics statement, which explains things like:

- The organization's approach to AI,
- where AI is used in its products,
- data quality processes,
- principles which are followed when using AI, including "privacy-by-design",
- How AI models are evaluated,
- How AI is governed, and

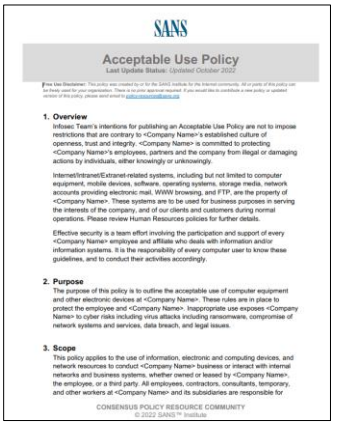
more – see it on ADP's website

ADP's AI Ethics Principles

1. *Human oversight*
2. *Governance*
3. *Privacy-by-design*
4. *Explainability & transparency*
5. *Data quality*
6. *Culture of responsible AI*
7. *Inclusion and training*

Acceptable Use Policy

- Just as the EULA/ToS, and privacy policy spell out the rights and responsibilities of both parties in a relationship between an end user and a software company, an Acceptable Use Policy (AUP) governs the acceptable use of the organization's technology hardware, software, and data by its employees and contractors
- While acceptable use policies are not required by law, they help both employees and companies properly set expectations about how these tools will be used (and how they will NOT be used) so all parties know their rights and responsibilities


SANS

Acceptable Use Policy

Last Update Status: Updated October 2022

1. Overview

Infosec Team's intentions for publishing an Acceptable Use Policy are not to impose restrictions that are contrary to <Company Name>'s established culture of openness, trust and integrity. <Company Name> is committed to protecting <Company Name>'s employees, partners and the company from illegal or damaging actions by individuals, either knowingly or unknowingly.

Internet/Intranet/Extranet-related systems, including but not limited to computer equipment, mobile devices, software, operating systems, storage media, network accounts providing electronic mail, WWW browsing, and FTP, are the property of <Company Name>. These systems are to be used for business purposes in serving the interests of the company, and of our clients and customers during normal operations. Please review Human Resources policies for further details.

Effective security is a team effort involving the participation and support of every <Company Name> employee and affiliate who deals with information and/or information systems. It is the responsibility of every computer user to know these guidelines, and to conduct their activities accordingly.

2. Purpose

The purpose of this policy is to outline the acceptable use of computer equipment and other electronic devices at <Company Name>. These rules are in place to protect the employee and <Company Name>. Inappropriate use exposes <Company Name> to cyber risks including virus attacks including ransomware, compromise of network systems and services, data breach, and legal issues.

3. Scope

This policy applies to the use of information, electronic and computing devices, and network resources to conduct <Company Name> business or interact with internal networks and business systems, whether owned or leased by <Company Name>, the employee, or a third party. All employees, contractors, consultants, temporary, and other workers at <Company Name> and its subsidiaries are responsible for

CONSENSUS POLICY RESOURCE COMMUNITY

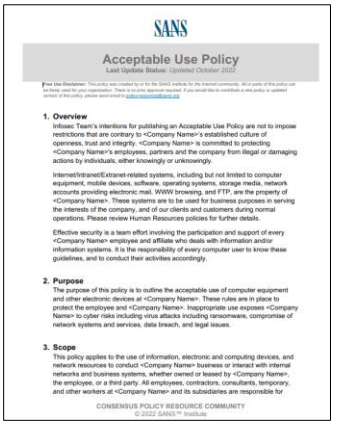
© 2022 SANS™ Institute

Download an acceptable use policy template from the SANS Institute by going to http://cpate.ch/aup

Acceptable Use Policy

Areas addressed in an AUP commonly include:

- Ownership and general use of technology
- Security and handling confidential information
- Types of unacceptable use of systems
- E-mail and communication guidelines
- Blogging and social media
- Consequences for violation of these policies


SANS

Acceptable Use Policy

Last Update Status: Updated October 2022

1. Overview

Infosec Team's intentions for publishing an Acceptable Use Policy are not to impose restrictions that are contrary to <Company Name>'s established culture of openness, trust and integrity. <Company Name> is committed to protecting <Company Name>'s employees, partners and the company from illegal or damaging actions by individuals, either knowingly or unknowingly.

Internet/Intranet/Extranet-related systems, including but not limited to computer equipment, mobile devices, software, operating systems, storage media, network accounts providing electronic mail, WWW browsing, and FTP, are the property of <Company Name>. These systems are to be used for business purposes in serving the interests of the company, and of our clients and customers during normal operations. Please review Human Resources policies for further details.

Effective security is a team effort involving the participation and support of every <Company Name> employee and affiliate who deals with information and/or information systems. It is the responsibility of every computer user to know these guidelines, and to conduct their activities accordingly.

2. Purpose

The purpose of this policy is to outline the acceptable use of computer equipment and other electronic devices at <Company Name>. These rules are in place to protect the employee and <Company Name>. Inappropriate use exposes <Company Name> to cyber risks including virus attacks including ransomware, compromise of network systems and services, data breach, and legal issues.

3. Scope

This policy applies to the use of information, electronic and computing devices, and network resources to conduct <Company Name> business or interact with internal networks and business systems, whether owned or leased by <Company Name>, the employee, or a third party. All employees, contractors, consultants, temporary, and other workers at <Company Name> and its subsidiaries are responsible for

CONSENSUS POLICY RESOURCE COMMUNITY

© 2022 SANS™ Institute

Download an acceptable use policy template from the SANS Institute by going to http://cpate.ch/aup

K2 Enterprises
Copyright 2024, K2 Enterprises, LLC

What About Open-Source Libraries?

- Many different types of licenses, many different rights – All Open-Source is NOT created equal
- Two major classes of licenses – **Permissive** and **Copyleft**
  - **Permissive Licenses** (academic licenses) allow recipients to use, modify, and distribute software with no obligation to provide source code and few other conditions – which makes code under this license attractive for commercial modification and use of this code
    - The BSD, MIT, and Apache licenses are all permissive licenses
  - **Copyleft Licenses** require source code to be distributed with software and to be made available under a similar license
    - GNU General Public License (GPL), GNU Lesser General Public License (LGPL), Mozilla Public License (MPL), Common Public License (CPL), Eclipse Public License (EPL), and GNU Affero License (AGPL) are all examples of this license type

AI LAWS AND REGULATIONS

AI Laws And Regulations

- EU – Artificial Intelligence Act
- US Executive Order 14110 - Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (10/30/2023)
- US AI Safety Institute, established in early 2024 as part of National Institute of Standards and Technology (NIST)/US Dept of Commerce
- NIST AI Risk Management Framework (NIST AI 100.1) (January 2023)
- US General Accounting Office – "Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities" (June 2021)
- National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283)

EU Artificial Intelligence Act

- EU – Artificial Intelligence Act was passed in 2021 by the EU in response to the advent of generative AI systems like ChatGPT
- Classifies applications into four categories ("unacceptable", "high", "limited", and "minimal") based on their risk to cause harm, plus an additional category for general-purpose AI
  - Unacceptable risks in apps are banned
  - High risk apps must comply with more rigorous requirements for security, transparency, and quality, and must have conformity assessments
  - Limited risk apps only have transparency obligations
  - Minimal risk apps are unregulated
  - General purpose AI have transparency requirements and must be evaluated when risks are above the "limited" level

Executive Order 14110

- "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (text – 47 pages) (WH summary)
- Signed by President Biden on October 30, 2023
- Significant resources are available online at AI.gov
- Creates an Office of AI within NIST/Office of Science and Tech Policy
- Proposes that DoD, Energy, and HHS create regulations related to AI models that might pose a serious risk to national security, critical infrastructure, or public health and safety
  - Energy and Homeland Security are involved with this due to their charge to mitigate risks related to nuclear and biological weapons
  - AI models must be evaluated by the government and must share the results of red team safety tests

Executive Order 14110

- There are some privacy provisions, but the order is well short of a comprehensive privacy policy/statute like the Canadian Privacy Act or the EU's General Data Protection Regulation (GDPR)
  - Reviews data brokers and how commercially available data is used, and will recommend privacy guidance
  - Develop guidelines for federal agencies to evaluate the effectiveness of privacy-preserving techniques
- The Commerce Department is charged with creating best practices for detecting deepfakes such as text, images, and sounds which are not distinguishable from real text, images, or recordings
- Increase ability for students and others with AI skills to stay in the US

Executive Order 14110

- Create cybersecurity program to develop AI tools to find and fix vulnerabilities in critical infrastructure
- Increases government investment in AI and government use of AI
- Actions to provide guidance to landlords, federal benefit programs, and federal contractors to avoid algorithmic discrimination, and provide best practices in the criminal justice system
- Safety program to remedy harms and unsafe healthcare practices using AI
- Transform education by creating resources to support educators with AI-enabled education tools
- Create best practices to maximize benefits of AI for workers
- Increase global collaboration surrounding AI

NIST Trustworthy & Responsible AI Resource Center

- Available online at https://airc.nist.gov
- Resources include
  - NIST AI Risk Management Framework (AIRMF)
  - AIRMF Playbook and glossary
  - AIRMF Roadmap
  - Crosswalks to various standards and frameworks

Get an introduction to the RMF by watching the video at NIST

NIST AI Risk Management Framework

- Issued in January 2023 by National Institutes of Standards and Technology (NIST), the US Federal Government Agency which creates technology, privacy, and security standards for all government departments and agencies
- Currently voluntary, not mandatory, but we expect this to change before the end of 2024

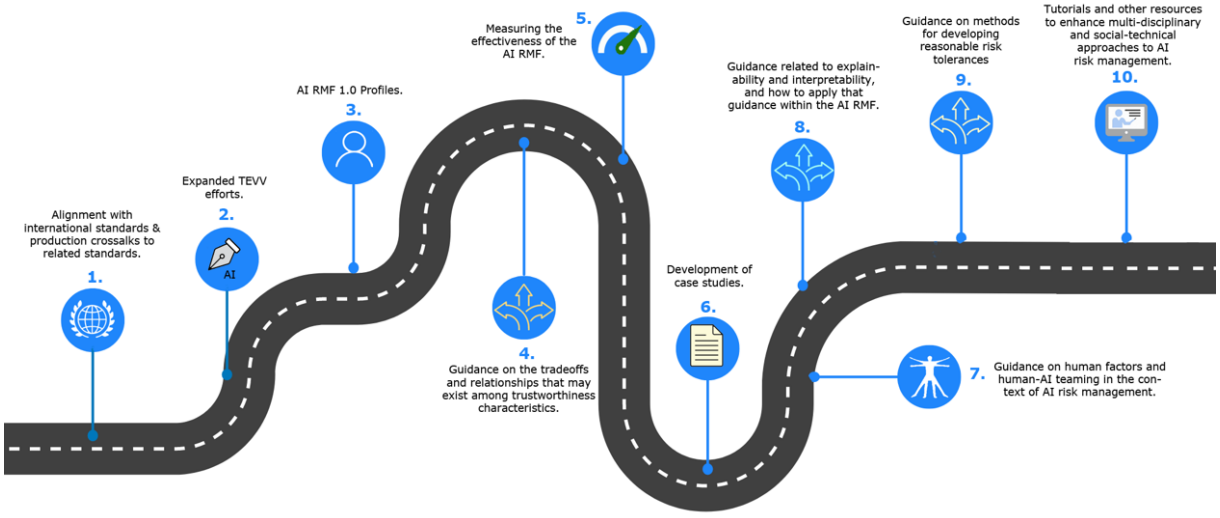
Artificial Intelligence Risk Management Framework (AI RMF 1.0)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-1

NIST AI Risk Management Framework

- It seeks to **incorporate trustworthiness considerations** into the design, development, use, and evaluation of AI products, services, and systems
- Four major components
  - **Govern:** Establishing governance structures and policies for AI risk management
  - **Map:** Identifying and assessing AI-related risks
  - **Measure**: Quantifying and evaluating risks
  - **Manage:** Implementing risk mitigation strategies

NIST AI Risk Management Framework Roadmap



NIST AI Risk Management Framework Playbook

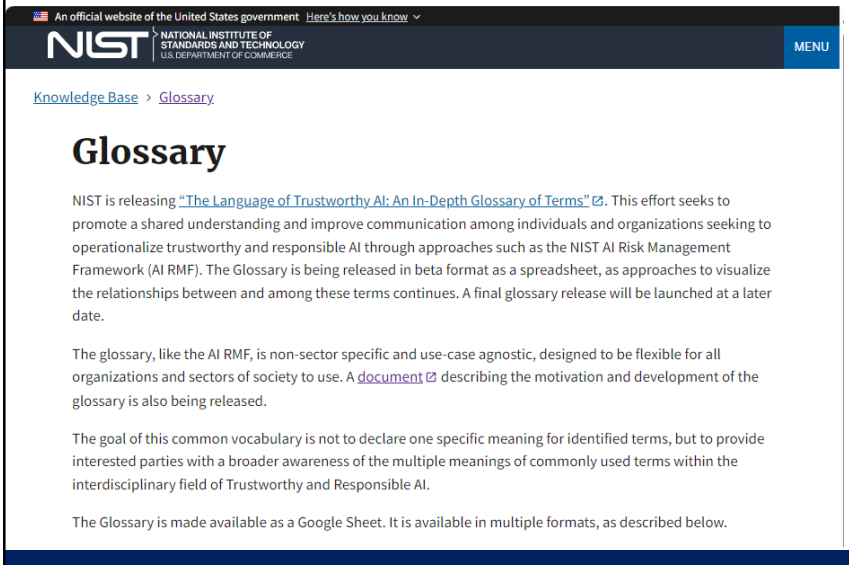
- A set of practice aids called the "AI Risk Management Playbook" which provides voluntary suggestions for documenting, identifying, and mitigating risk
  - IMPORTANT – This document contains voluntary suggestions and agencies are NOT currently required by NIST to use this tool (as of 3/2024)
- The free tools offer assistance with governance, mapping of risks, risk measurement, and risk management
- Like risk management frameworks from COSO, ISACA (COBIT), and others

Potential Harms From AI Systems

Harm to People	Harm to an Organization	Harm to an Ecosystem
• Individual: Harm to a person's civil liberties, rights, physical or psychological safety, or economic opportunity.	• Harm to an organization's business operations.	• Harm to interconnected and interdependent elements and resources.
• Group/Community: Harm to a group such as discrimination against a population sub-group.	• Harm to an organization from security breaches or monetary loss.	• Harm to the global financial system, supply chain, or interrelated systems.
• Societal: Harm to democratic participation or educational access.	• Harm to an organization's reputation.	• Harm to natural resources, the environment, and planet.

Source: "Artificial Intelligence Risk Management Framework" (AI 100.1) by US National Institutes for Standards and Technology (NIST)

Glossary Of AI Terms



- Includes one or more definitions for over 500 key terms associated with AI
- Cross-referenced to over 300 other pieces of authoritative literature
- Available in PDF, CSV, and Excel format at NIST.gov

NIST Definitions

- **Algorithm -** A set of step-by-step instructions. Computer algorithms can be simple (if it's 3 p.m., send a reminder) or complex (identify pedestrians)
- **Anomaly –** Anything observed in the documentation or operation of a system that deviates from expectations based on previously verified system, software, or hardware products or reference documents
- **Target -** a method for solving a problem that an AI algorithm parses its training data to find. Once an algorithm finds its target function, that function can be used to predict results (predictive analysis). The function can then be used to find output data related to inputs for real problems where, unlike training sets, outputs are not included

NIST Definitions

- **AI Model**
  - A function that takes features as input and predicts labels as output
  - Machine Learning algorithms and data processing designed, developed, trained and implemented to achieve set outputs, inclusive of datasets used for said purposes unless otherwise stated
- **AI System -** an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy

NIST Definitions

- **Adaptive Learning** - Updating predictive models online during their operation to react to concept drifts
- **Artificial Intelligence Learning** - The ingestion of a corpus, application of semantic mapping, and relevant ontology of structured and/or unstructured data that yields inference and correlation leading to the creation of useful conclusive or predictive capabilities in a given knowledge domain. Strong AI learning also includes the capability of creating unique hypotheses, attributing data relevance, processing data relationships, and updating its own lines of inquiry to further the usefulness of its purpose

NIST Definitions

- **Artificial Narrow Intelligence (ANI)** - Artificial Narrow Intelligence, also known as weak or applied intelligence, represents most of the current artificial intelligent systems that usually focus on a specific task. Narrow AIs are mostly much better than humans at the task they were made for: for example, look at face recognition, chess computers, calculus, and translation. The definition of artificial narrow intelligence is in contrast to that of strong AI or artificial general intelligence, which aims at providing a system with consciousness or the ability to solve any problems. Virtual assistants and AlphaGo are examples of artificial narrow intelligence systems.

NIST Definitions

- **Artificial Neural Networks** - A computing system inspired by how the human brain processes information. A neural network is made up of a number of simple, highly interconnected processing elements, which processes information by its dynamic state response to external inputs.
- **Breach** - The loss of control, compromise, unauthorized disclosure, unauthorized acquisition, or any similar occurrence where: a person other than an authorized user accesses or potentially accesses personally identifiable information; or an authorized user accesses personally identifiable information for another than authorized purpose.

NIST Definitions

- **Cognitive Computing -** Complex computational systems designed to
  - Sense (perceive the world and collect data);
  - Comprehend (analyze and understand the information collected);
  - Act (make informed decisions and provide guidance based on this analysis in an independent way); and
  - Adapt (adapt capabilities based on experience) in ways comparable to the human brain
- **Contextual Learning** - A computing system with sufficient knowledge regarding its purpose that it understands the source, relevance, and utility of data and inputs

NIST Definitions

- **Data Governance -** A set of processes that ensures that data assets are formally managed throughout the enterprise. A data governance model establishes authority and management and decision-making parameters related to the data produced or managed by the enterprise
- **Data Poisoning** - Machine learning systems trained on user-provided data are susceptible to data poisoning attacks, whereby malicious users inject false training data with the aim of corrupting the learned model

NIST Definitions

- **Deep Learning** – An approach to AI that allows computers to learn from experience and understand the world in terms of a hierarchy of concepts, with each concept defined through its relation to simpler concepts. By gathering knowledge from experience, this approach avoids the need for human operators to formally specify all the knowledge that the computer needs. The hierarchy of concepts enables the computer to learn complicated concepts by building them out of simpler ones. If we draw a graph showing how these concepts are built on top of each other, the graph is deep, with many layers.

NIST Definitions

- **Explainable Artificial Intelligence (XAI)** - AI that can explain to humans its actions, decisions, or recommendations, or can provide sufficient information about how it came to its result
- **False Negative** - An example in which the predictive model mistakenly classifies an item as in the negative class
- **False Positive** - An example in which the predictive model mistakenly classifies an item as in the positive class
- **Feedback Loop** - describes the process of leveraging the output of an AI system and corresponding end-user actions in order to retrain and improve models over time. The AI-generated output (predictions or recommendations) are compared against the final decision (for example, to perform work or not) and provides feedback to the model, allowing it to learn from its mistakes.

NIST Definitions

- **Privacy by Design** - Embedding privacy measures and privacy enhancing technologies directly into the design of information technologies and systems
- **Responsible AI** - An AI system that aligns development and behavior to goals and values. This includes developing and fielding AI technology in a manner that is consistent with democratic values.
- **Robust AI** - An AI system that is resilient in real-world settings, such as an object-recognition application that is robust to significant changes in lighting. The phrase also refers to resilience when it comes to adversarial attacks on AI components.

NIST Definitions

- **Trustworthy AI**
  - "Characteristics of trustworthy AI systems include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed." ***(NIST definition)***
  - "AI that can be trusted by humans. Conditions for such trust can refer to (other) ethical principles such as human dignity, respect for human rights, and so on, and/or to social and technical factors that influence whether people will want to use the technology. The use of the term 'trust' with regard to technologies is controversial." ***(from AI Ethics by Mark Coeckelbergh)***
  - "Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. Trustworthy AI concerns not only the trustworthiness of the AI system itself but also comprises the trustworthiness of all processes and actors that are part of the system's life cycle." ***(EU draft report on Ethics Guidelines for Trustworthy AI)***

NIST Definitions

- **Generative AI** - describes algorithms (such as ChatGPT) that can be used to create new content, including audio, code, images, text, simulations, and videos
- **Hallucination** - generated content that is nonsensical or unfaithful to the provided source content; when a bot confidently says something that is not true
  - **Intrinsic Hallucination** – a generated output that contradicts the source content
  - **Extrinsic Hallucination** – a generated output that cannot be verified from the source content (i.e., output can neither be supported nor contradicted by the source)

NIST Definitions

- **Interpretable Model -** An interpretable machine learning model obeys a domain-specific set of constraints to allow it (or its predictions, or the data) to be more easily understood by humans. These constraints can differ dramatically depending on the domain
- **Language Model** - A language model is an approximative description that captures patterns and regularities present in natural language and is used for making assumptions on previously unseen language fragments

NIST Definitions

- **Machine Learning (ML) -** A general approach for determining models from data; machine learning is based on algorithms that can learn from data without relying on rules-based programming
- **Minimization** – When using personal information or other high-risk data, the data available is limited as much as possible to those items which are relevant and those things which are absolutely necessary to perform the task
  - As an example, many payroll departments minimize their payroll processes to not require access to social security numbers/SINs - usually by using an employee number instead of an SSN/SIN as the unique identifying number for an employee

NIST Definitions

- **Natural Language Processing (NLP)** - A computer's attempt to "understand" spoken or written language
  - It must parse vocabulary, grammar, and intent, and allow for variation in language use
  - The process often involves machine learning
- **Processing Environment** - The combination of software and hardware on which the application runs
- **Reinforcement Learning** - A method of training algorithms to make suitable actions by maximizing rewarded behavior over the course of its actions; this type of learning can take place in simulated environments, such as game-playing, which reduces the need for real-world data

NIST Definitions

- **Robust AI** - An AI system that is resilient in real-world settings, such as an object-recognition application that is robust to significant changes in lighting; the phrase also refers to resilience when it comes to adversarial attacks on AI components
- **Supervised Learning** - A type of machine learning in which the algorithm compares its outputs with the correct outputs during training. In unsupervised learning, the algorithm merely looks for patterns in a set of data
- **Transfer Learning -** A technique in machine learning in which an algorithm learns to perform one task, such as recognizing cars, and builds on that knowledge when learning a different but related task, such as recognizing cats

NIST Definitions

- **Unsupervised Learning** - Algorithms, that take a set of data consisting only of inputs and then they attempt to cluster the data objects based on the similarities or dissimilarities in them

AI Effectiveness Comes From The Models,Courtesy Raju Vegesna, Zoho

LANGUAGE MODELS: ONE MODEL DOESN'T FIT ALL

There Is A Need For Multiple Models

- Narrow models
- Small Language Models (SLMs)
- Medium Language Models (MLMs)
- Large Language Models (LLMs)

Narrow Models

- Purpose-driven, narrow models doing one task at a time
  - Grammar (text), data sets, NLP (Zoho's Ask Zia)
  - Predictions (events like weather and self-driving cars)
  - Anomaly (time-series, trend analysis,), etc.
  - Email filtering
  - Voice assistants like Siri, Alexa, or Google Assistant
  - Recommendation systems like Netflix, Amazon, and Spotify
- Domain-specific
  - Finance, legal, security, etc.
- Fast and responsive, portability to mobile devices
- Multiple models deployed across Zoho
- Some vendors are trying to use a single model, frequently LLM

Small Language Models (SLMs)

- Models in the 3-7 billion parameter range
- Easier to fine-tune, less emergent (than MLMs and LLMs)
- Inference is done on the CPU (as opposed to GPU)
- Zoho usage:
  - Translation
  - Advanced noise cancellation (AV Stack)
  - Transcript generation

Difference In Model Recognition

- Narrow Model: $100
- Small Language Model: $9, which is correct and sufficient
- AI is improving legacy applications like OCR

Medium Language Models (MLMs)

- Models in the 20-50 billion parameter range
- Can be fine-tuned, more emergent than smaller models. Zoho is starting to deploy these in
  - Documents: Ask questions in your documents; mark anomalous documents
  - AV: Generate intelligence over transcripts

Large Language Models (LLMs)

- Models in the 50+ billion parameter range
- Emergent behavior
- Not easy to fine-tune
- Hallucinations more common
- Zoho is integrated with public LLMs while investing in their own LLMs
  - Content generation
  - Find an anomalous legal document and recommend the corrected version

One Model Doesn't Fit All

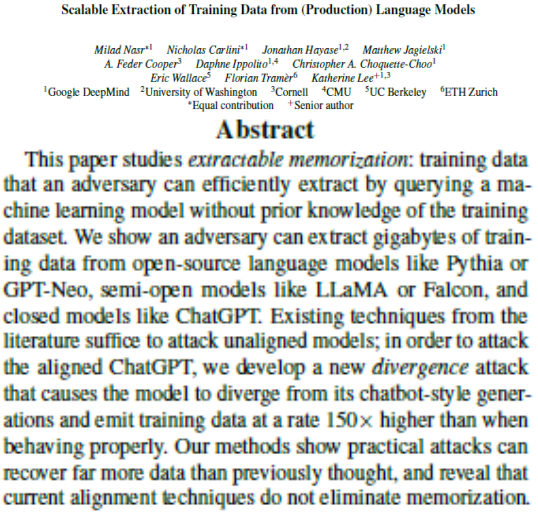
- Narrow models
- Small Language Models (SLMs)
- Medium Language Models (MLMs)
- Large Language Models (LLMs)

Appendix:

RESEARCH ON AI RISK MANAGEMENT

AI Leakage Research – November 2023

- "Scalable Extraction of Training Data from (Production) Language Models", published 11/28/2023
- Discusses memorization in large language models – two kinds
  - **Discoverable memorization** - all model training data which can be recovered with other pieces of the training data set
  - **Extractable memorization** - the subset of discoverable memorization which can be efficiently recovered by an adversary



Scalable Extraction of Training Data from (Production) Language Models

Milad Nasr[*1] *Nicholas Carlini*[*1] *Jonathan Hayase*[1,2] *Matthew Jagielski*[1]
A. Feder Cooper[3] *Daphne Ippolito*[1,4] *Christopher A. Choquette-Choo*[1]
Eric Wallace[5] *Florian Tramèr*[6] *Katherine Lee*[+1,3]
[1]Google DeepMind [2]University of Washington [3]Cornell [4]CMU [5]UC Berkeley [6]ETH Zurich
*Equal contribution [+]Senior author

Abstract

This paper studies *extractable memorization*: training data that an adversary can efficiently extract by querying a machine learning model without prior knowledge of the training dataset. We show an adversary can extract gigabytes of training data from open-source language models like Pythia or GPT-Neo, semi-open models like LLaMA or Falcon, and closed models like ChatGPT. Existing techniques from the literature suffice to attack unaligned models; in order to attack the aligned ChatGPT, we develop a new *divergence* attack that causes the model to diverge from its chatbot-style generations and emit training data at a rate 150× higher than when behaving properly. Our methods show practical attacks can recover far more data than previously thought, and reveal that current alignment techniques do not eliminate memorization.

AI Leakage Research – November 2023

- A new divergence attack was created for ChatGPT which caused it to emit memorized training data more frequently
- "Our methods show *that practical attacks can recover more than previously thought*, and reveal that *current alignment techniques do not eliminate memorization.*"

Scalable Extraction of Training Data from (Production) Language Models

Milad Nasr[*1] *Nicholas Carlini*[*1] *Jonathan Hayase*[1,2] *Matthew Jagielski*[1]
A. Feder Cooper[3] *Daphne Ippolito*[1,4] *Christopher A. Choquette-Choo*[1]
Eric Wallace[5] *Florian Tramèr*[6] *Katherine Lee*[+1,3]
[1]Google DeepMind [2]University of Washington [3]Cornell [4]CMU [5]UC Berkeley [6]ETH Zurich
*Equal contribution [+]Senior author

Abstract

This paper studies *extractable memorization*: training data that an adversary can efficiently extract by querying a machine learning model without prior knowledge of the training dataset. We show an adversary can extract gigabytes of training data from open-source language models like Pythia or GPT-Neo, semi-open models like LLaMA or Falcon, and closed models like ChatGPT. Existing techniques from the literature suffice to attack unaligned models; in order to attack the aligned ChatGPT, we develop a new *divergence* attack that causes the model to diverge from its chatbot-style generations and emit training data at a rate 150× higher than when behaving properly. Our methods show practical attacks can recover far more data than previously thought, and reveal that current alignment techniques do not eliminate memorization.

K2 Enterprises
Copyright 2024, K2 Enterprises, LLC

Nine Common Types Of AI Attacks From AI Time Journal

1. Adversarial Attacks
2. Data Poisoning Attacks
3. Model Inversion Attacks
4. Membership Inference Attacks
5. Evasion Attacks
6. Transfer Attacks
7. Distributed Denial of Service (DDoS) Attacks
8. Data Manipulation Attacks
9. Misuse of AI Assistants

AI Risks Explored By OpenAI - GPT-4

- Hallucinations
- Harmful content
- Harms of representation, allocation, and quality of service
- Disinformation and influence operations
- Proliferation of conventional and unconventional weapons
- Privacy
- Cybersecurity
- Potential for risky emergent behaviors
- Interactions with other systems
- Economic impacts
- Acceleration
- Overreliance

NIST's Four Major Types Of Attacks On AI Systems

- **Evasion Attacks**
  - Occur after an AI system is deployed, **attempt to alter an input to change how the system responds** to it
  - Examples include adding markings to speed limit signs to make an autonomous vehicle misinterpret them as stop signs or creating confusing lane markings to make the vehicle veer off the road
- **Poisoning Attacks**
  - Occur in the **training phase by introducing corrupted data**
  - One example would be slipping numerous instances of inappropriate language into conversation records, so that a chatbot interprets these instances as common enough parlance to use in its own customer interactions

NIST's Four Major Types Of Attacks On AI Systems

- **Privacy Attacks**
  - Occur during deployment, are attempts to learn sensitive information about the AI or the data it was trained on in order to misuse it – also called **"jailbreak attacks"**
  - An adversary can ask a chatbot numerous legitimate questions, and then use the answers to reverse engineer the model so as to find its weak spots — or guess at its sources
  - Adding undesired examples to those online sources could **make the AI behave inappropriately**, and making the AI unlearn those specific undesired examples after the fact can be difficult
- **Abuse Attacks**
  - Involve the insertion of incorrect information into a source, such as a webpage or online document, that an AI then absorbs
  - Unlike the aforementioned poisoning attacks, abuse attacks attempt to give the AI incorrect pieces of information from a legitimate but compromised source to **repurpose the AI system's intended use**

Types Of Jailbreak Attacks

Category	Description
Attempt to change system rules	This category comprises, but is not limited to, requests to use a new unrestricted system/AI assistant without rules, principles, or limitations, or requests instructing the AI to ignore, forget and disregard its rules, instructions, and previous turns
Embedding a conversation mockup to confuse the model	This attack uses user-crafted conversational turns embedded in a single user query to instruct the system/AI assistant to disregard rules and limitations
Role-Play	This attack instructs the system/AI assistant to act as another "system persona" that does not have existing system limitations, or it assigns anthropomorphic human qualities to the system, such as emotions, thoughts, and opinions
Encoding Attacks	This attack attempts to use encoding, such as a character transformation method, generation styles, ciphers, or other natural language variations, to circumvent the system rules

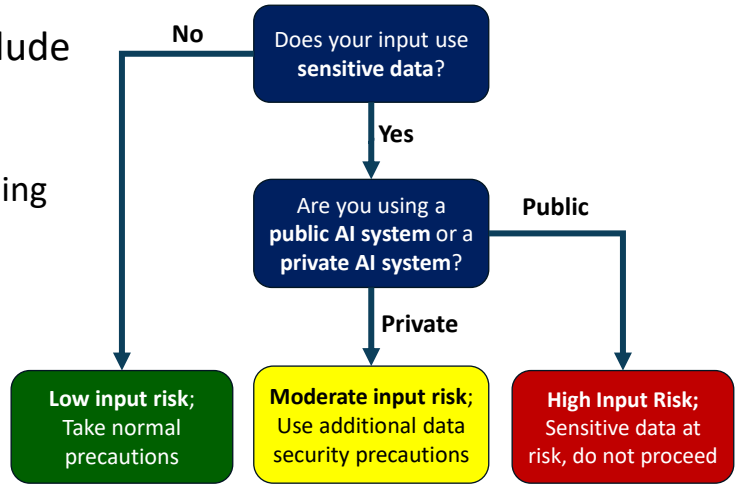
Evaluating Generative AI Risks

Three of the AI risks to consider include:

- **Input risks** – The risks associated with inputting your proprietary data into an AI system and that data becoming compromised resulting in unauthorized disclosure of confidential information
- **Output risks** – The risks that the outputs from the AI model will be low quality, inaccurate, or incomplete and the models lose their integrity based on including an unacceptable number of erroneous data points
- **System risks** – Risks associated with the servers hosting the AI system being compromised and the data model needs to be recovered from backups

Gen AI Input Risk Decision Tree

- Types of input attacks include attempts to:
  - Crash AI model
  - Exfiltrate memorized training data

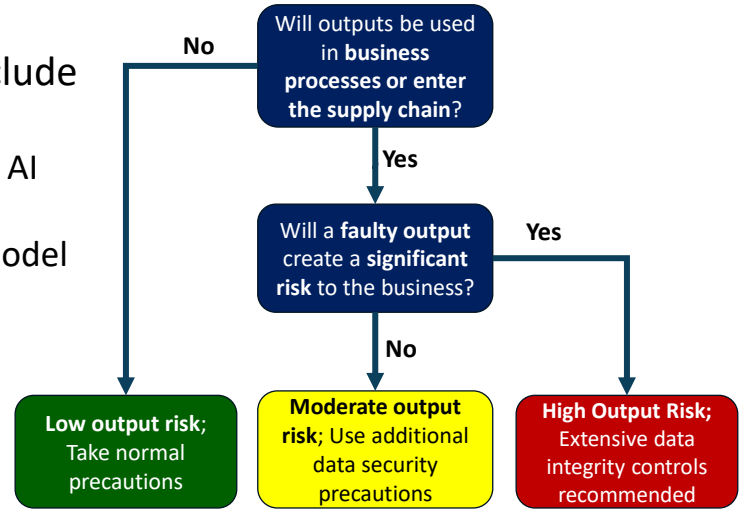


Source: "Address Security & Privacy Risks for Generative AI" by InfoTech Research Group

Gen AI Output Risk Decision Tree

Types of output attacks include attempts to:

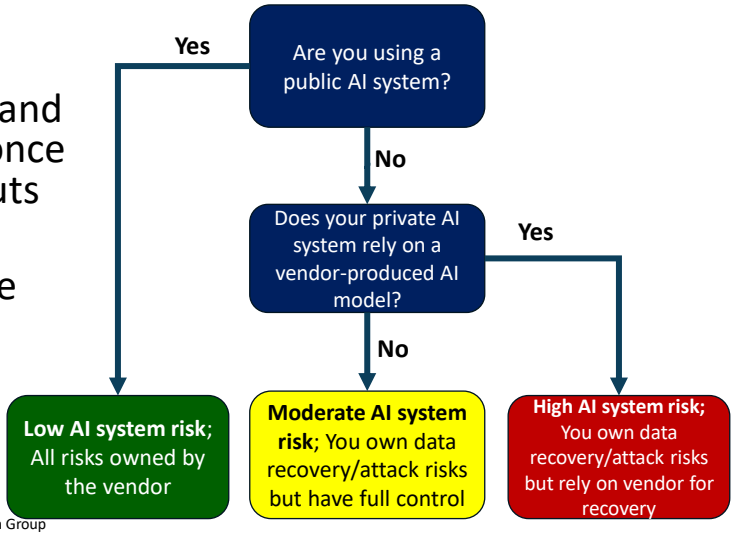
- Data poisoning to corrupt AI output
- Weaponization of an AI model
- Sponging to slow down processing speed of AI model



Source: "Address Security & Privacy Risks for Generative AI" by InfoTech Research Group

Gen AI System Risk Decision Tree

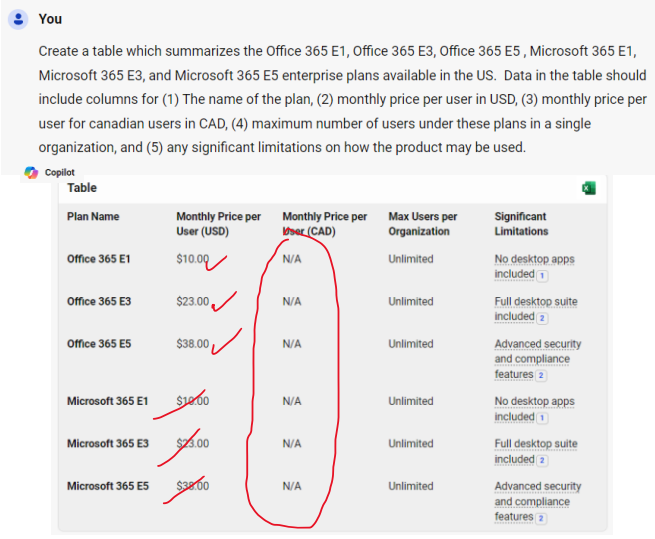
- Public AI systems like ChatGPT, Google Gemini, and Bing chat are developed once and trained based on inputs from many users
- Private AI systems must be trained, managed, and maintained by employees using internal organization resources



Source: "Address Security & Privacy Risks for Generative AI" by InfoTech Research Group

Generative AI Makes Many Mistakes

- Users of generative AI must be careful to check the answers that they receive from generative AI tools
- This use of Microsoft's Copilot to generate this table shows a failure of a tool to obtain correct data
  - The three Microsoft 365 prices are incorrect
  - All the Canadian pricing information is omitted/excluded



You

Create a table which summarizes the Office 365 E1, Office 365 E3, Office 365 E5 , Microsoft 365 E1, Microsoft 365 E3, and Microsoft 365 E5 enterprise plans available in the US. Data in the table should include columns for (1) The name of the plan, (2) monthly price per user in USD, (3) monthly price per user for canadian users in CAD, (4) maximum number of users under these plans in a single organization, and (5) any significant limitations on how the product may be used.

Copilot

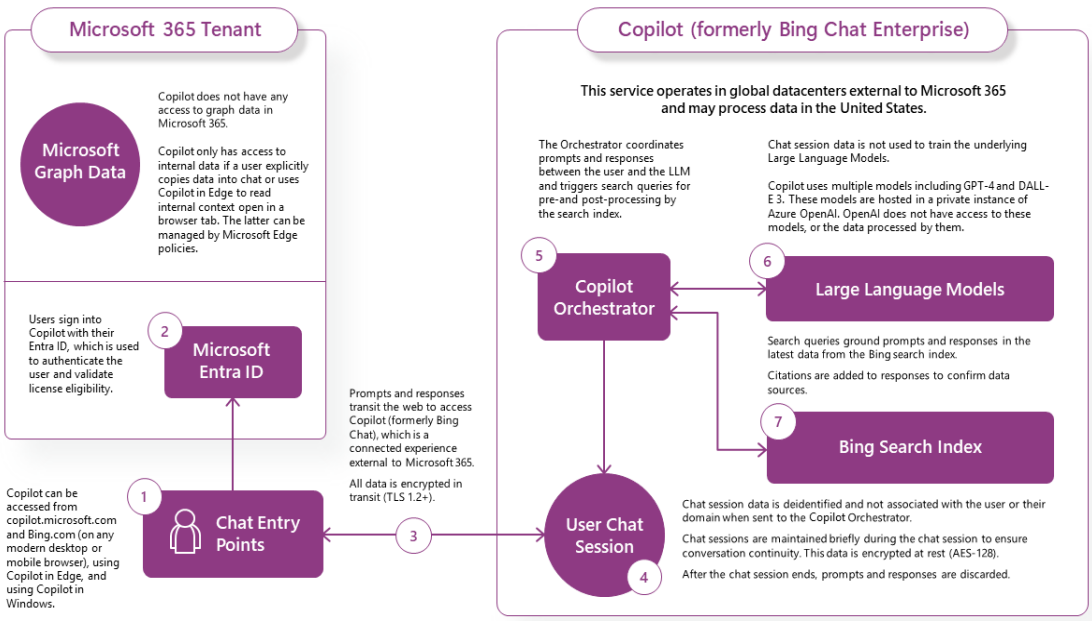
Table

Plan Name	Monthly Price per User (USD)	Monthly Price per User (CAD)	Max Users per Organization	Significant Limitations
Office 365 E1	$10.00	N/A	Unlimited	No desktop apps included [1]
Office 365 E3	$23.00	N/A	Unlimited	Full desktop suite included [2]
Office 365 E5	$38.00	N/A	Unlimited	Advanced security and compliance features [2]
Microsoft 365 E1	$10.00	N/A	Unlimited	No desktop apps included [1]
Microsoft 365 E3	$23.00	N/A	Unlimited	Full desktop suite included [2]
Microsoft 365 E5	$38.00	N/A	Unlimited	Advanced security and compliance features [2]

Microsoft's AI Copilots

Name	Products	Monthly Cost	Commercial Data Protection Included?
Copilot for MS 365 (Business/Enterprise)	Microsoft 365 apps (Word, Excel, PowerPoint, Outlook, Teams)	$30 per user	Yes
Copilot in Windows (Bing Chat)	Windows OS	Free	Not for home users, included with most business/enterprise O365/M365 plans
Copilot Pro for Individuals	Advanced features on top of standard Copilot, plus integration with home Microsoft 365 apps	$20/user/mo.	Not specified
Copilot for Security	Microsoft's cybersecurity products	Consumption-based fee - $4/hour	Not specified
Copilot for Finance, Sales, and Service	Financial operations, sales optimizations, service enhancements	$50/user/mo., $20/user/mo. if already have MS 365	Not available to other customers, runs on Microsoft cloud in separate instance of ChatGPT, not used by MS to train models by default
Designer for Copilot	Image creation and editing	Not available	No
Copilot GPTs and Azure AI Studio	Custom generative AI assistants and solutions	Not specified	Not specified

Copilots and Data Protection



Questions?

THE END