

Wydział Finansów i Bankowości

Filip Kasprzyk

**Generative AI – Ryzyka oraz Cyberbezpieczeństwo**

Praca magisterska

**Promotor**

**dr Mariusz Nogala**

Poznań 2024

**Spis treści**

[1. Wstęp 3](#_Toc166608371)

[2 Wprowadzenie do Generatywnego AI. 4](#_Toc166608372)

[2.1 Definicja Generatywnego AI 4](#_Toc166608373)

[2.2 Przegląd technologii i algorytmów Generatywnego AI 5](#_Toc166608374)

[2.2.1 Generative Adversarial Networks (GANs) 5](#_Toc166608375)

[2.2.2 Variational Autoencoders (VAEs) 6](#_Toc166608376)

[2.2.3 Transformer Models 8](#_Toc166608377)

[2.3 Zastosowania Generatywnego AI 11](#_Toc166608378)

[2.3.1 Wykrywanie i tworzenie ataków phishingowych 11](#_Toc166608379)

[2.3.2 Maskowanie danych i ochrona prywatności 13](#_Toc166608380)

[2.3.3 Odpowiedź na incydenty 15](#_Toc166608381)

[3 Ryzyka w Generatywnym AI. 17](#_Toc166608382)

[4 Bezpieczeństwo w Generatywnym AI. 17](#_Toc166608383)

[5 Rozdział 4. 17](#_Toc166608384)

[6 Rozdział 5. 17](#_Toc166608385)

[7 Bibliografia 17](#_Toc166608386)

# **Wstęp**

W dobie cyfryzacji i rosnącej zależności od technologii, sztuczna inteligencja (AI) stała się nieodłącznym elementem naszego codziennego życia. Wśród różnych dziedzin AI, **Generative AI**, czyli generatywna sztuczna inteligencja, zyskuje na znaczeniu ze względu na swoją zdolność do tworzenia nowych, wcześniej niewidzianych danych na podstawie nauczonych wzorców.

Jednak, jak każda technologia, Generative AI nie jest wolna od ryzyka. Wraz z niesamowitymi możliwościami, które oferuje, pojawiają się również nowe wyzwania i zagrożenia dla cyberbezpieczeństwa. Od fałszywych informacji i deepfake’ów, po potencjalne naruszenia prywatności i bezpieczeństwa danych, ryzyka związane z Generative AI są różnorodne i złożone.

Ta praca magisterska ma na celu zbadanie tych ryzyk i wyzwań związanych z Generative AI w kontekście cyberbezpieczeństwa. Przeanalizujemy różne scenariusze, w których Generative AI może być wykorzystywane do celów złośliwych, a także omówimy strategie i rozwiązania, które mogą pomóc w minimalizacji tych ryzyk.

Celem tej pracy jest nie tylko zrozumienie potencjalnych zagrożeń, ale także podkreślenie znaczenia etyki i odpowiedzialności w rozwoju i zastosowaniu Generative AI. Wierze, że zrozumienie i adresowanie tych kwestii jest kluczowe dla bezpiecznego i odpowiedzialnego wykorzystania tej fascynującej technologii.

# **Wprowadzenie do Generatywnego AI.**

## **Definicja Generatywnego AI**

Generatywne AI, znane również jako GenAI, to dziedzina sztucznej inteligencji, która koncentruje się na tworzeniu nowych, wcześniej nieistniejących danych, które mogą być nieodróżnialne od danych, które mogłyby zostać wygenerowane przez człowieka. (1)(2)(3)

GenAI wykorzystuje różne techniki, takie jak Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) i Transformer Models, aby generować dane, które mogą obejmować tekst, obrazy, muzykę, dźwięk, wideo, a nawet 3D renderings.

W kontekście cyberbezpieczeństwa, GenAI reprezentuje transformacyjną zmianę w sposobie, w jaki profesjonaliści ds. bezpieczeństwa przewidują, wykrywają i reagują na zagrożenia123. Ta technologia wykorzystuje modele uczenia maszynowego, szczególnie te oparte na GANs, do symulowania ataków cybernetycznych i strategii obronnych. (1)(2)(3)

Możliwość generowania nowych instancji danych, które naśladują rzeczywiste zestawy danych, pozwala systemom cyberbezpieczeństwa szybko ewoluować, dostosowując się do nowych zagrożeń, gdy się pojawiają123. W miarę jak te modele AI przechodzą szkolenie, stają się coraz bardziej zaawansowane w zrozumieniu niuansów danych bezpieczeństwa, co pozwala im identyfikować subtelne wzorce działalności złośliwej, które mogą umknąć tradycyjnym metodom wykrywania. (1)(2)(3)

Jednak, jak każda potężna technologia, GenAI niesie ze sobą również swoje wyzwania i ryzyka. Na przykład, GenAI może być wykorzystywane do celów złośliwych, takich jak tworzenie deepfakes, które mogą być używane do manipulacji i oszustw123. Ponadto, ponieważ GenAI jest trenowane na dużych zestawach danych, istnieje ryzyko naruszenia prywatności, jeśli dane te zawierają informacje osobiste123. Wreszcie, istnieje ryzyko, że GenAI może generować informacje, które są oczywiście fałszywe, znanymi jako “halucynacje”, co może prowadzić do dezinformacji i nieporozumień. (1)(2)(3)

W związku z tym, istotne jest, aby podejść do GenAI z solidnym podejściem zero-trust, aby uniknąć niepożądanej ingerencji w wyniki funkcji GenAI3. Wymaga to moralnego monitoringu i ciągłego rozwoju, aby rozwiązać te problemy, zapewnić zgodność tych modeli z naszymi wartościami moralnymi i zredukować błędy. (1)(2)(3)

## **Przegląd technologii i algorytmów Generatywnego AI**

Generatywne AI to fascynująca dziedzina, która wykorzystuje różne technologie i algorytmy do tworzenia nowych, wcześniej nieistniejących danych. W tym rozdziale przyjrzymy się bliżej niektórym z tych technologii i algorytmów, które napędzają GenAI, w tym Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) i Transformer Models.

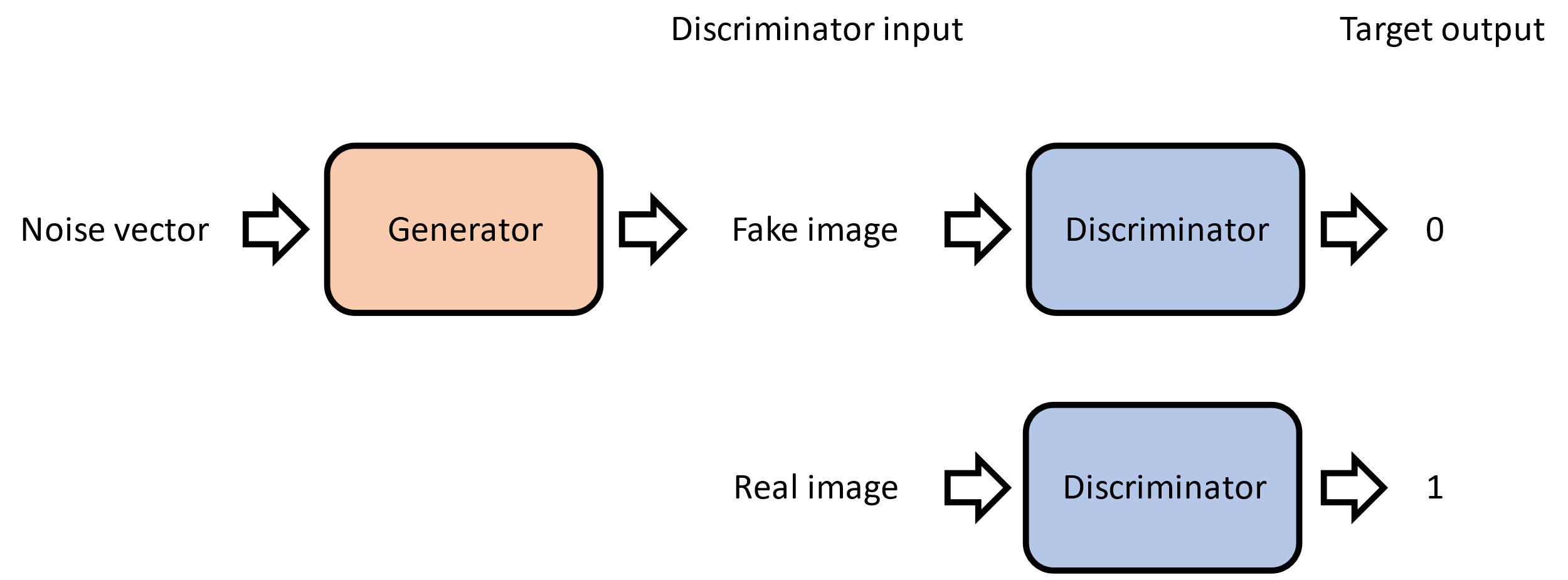
Każda z tych technologii ma swoje unikalne cechy i zastosowania, a razem tworzą potężne narzędzie, które może przekształcić wiele dziedzin, od sztuki po naukę. Przyjrzymy się, jak te technologie działają, jakie są ich mocne i słabe strony, oraz jak są wykorzystywane w praktyce.

Zanim jednak zagłębimy się w szczegóły, warto zrozumieć, dlaczego te technologie są tak ważne. Generatywne AI otwiera nowe możliwości dla twórczości i innowacji, umożliwiając tworzenie rzeczy, które wcześniej były niemożliwe do wyobrażenia. Jednak jak każda potężna technologia, GenAI niesie ze sobą również swoje wyzwania i ryzyka, które musimy zrozumieć i zarządzać. W kolejnych podrozdziałach przyjrzymy się bliżej wspomnianych wyżej technologii

### **Generative Adversarial Networks (GANs)**

**Generatywne sieci przeciwnościowe**, znane jako **GANs**, to rodzaj sieci neuronowej, która składa się z dwóch części: generatora i dyskryminatora. Generator tworzy nowe dane, a dyskryminator próbuje odróżnić te wygenerowane dane od prawdziwych. W wyniku tej rywalizacji, generator uczy się tworzyć coraz bardziej przekonujące dane. (4)

**GANs** są często używane do generowania obrazów, ale mogą być również używane do generowania innych rodzajów danych123. Są one wykorzystywane w wielu dziedzinach, od sztuki po naukę, i mają ogromny potencjał, ale także stwarzają wyzwania związane z bezpieczeństwem i etyką. (4)

**GANs** są podobne do naśladownictwa w biologii ewolucyjnej, z ewolucyjnym wyścigiem zbrojeń między obiema sieciami. Generator stara się zminimalizować cel, a dyskryminator dąży do maksymalizacji celu. Zadaniem generatora jest zbliżenie się do , to znaczy dopasowanie swojego własnego rozkładu wyjściowego jak najbliżej do rozkładu referencyjnego2. Zadaniem dyskryminatora jest wyjście wartości bliskiej 1, gdy wejście wydaje się pochodzić z rozkładu referencyjnego, i wyjście wartości bliskiej 0, gdy wejście wygląda na pochodzące z rozkładu generatora. (4)

Rysunek 1 - Schemat działania GANs

Źródło: Generative adversarial network. <https://en.wikipedia.org/wiki/Generative_adversarial_network>

### **Variational Autoencoders (VAEs)**

Variational Autoencoders (VAEs) to rodzaj sieci neuronowej, która uczy się reprezentować dane w ukrytej przestrzeni, a następnie generować nowe dane w tej przestrzeni. VAEs są często używane do generowania obrazów, ale mogą być również używane do generowania innych rodzajów danych. (5)(6)

**VAEs** są potężnymi modelami generatywnymi, które uczą się probabilistycznej reprezentacji danych za pomocą technik uczenia głębokiego. Składają się z sieci kodera i dekodera, które mapują punkty danych między oryginalną przestrzenią danych a przestrzenią o niższym wymiarze.

Koder w **VAE** ma na celu nauczenie się efektywnego kodowania danych z zestawu danych i przekazanie go do architektury typu "bottleneck". Inna część autoenkodera to dekoder, który korzysta z przestrzeni ukrytej w warstwie “bottleneck” do regeneracji obrazów podobnych do zestawu danych.

**VAE** różni się od autoenkodera tym, że dostarcza statystycznego sposobu opisywania próbek zestawu danych w przestrzeni ukrytej. Dlatego w **VAE**, koder generuje rozkład prawdopodobieństwa w warstwie “bottleneck” zamiast pojedynczej wartości wyjściowej.

Architektura koder-dekoder leży u podstaw Variational Autoencoders (**VAEs**), odróżniając je od tradycyjnych autoenkoderów5. Sieć kodera przekształca surowe dane wejściowe i przekształca je w rozkład prawdopodobieństwa w przestrzeni ukrytej. Kod ukryty generowany przez koder jest kodowaniem probabilistycznym, co pozwala VAE wyrazić nie tylko pojedynczy punkt w przestrzeni ukrytej, ale także rozkład potencjalnych reprezentacji.

Sieć dekodera z kolei bierze próbkowany punkt z rozkładu ukrytego i odtwarza go z powrotem do przestrzeni danych. Podczas treningu, model doskonali zarówno parametry kodera, jak i dekodera, aby zminimalizować stratę rekonstrukcji - różnicę między danymi wejściowymi a wyjściem dekodowanym. Celem nie jest tylko osiągnięcie dokładnej rekonstrukcji, ale także regulacja przestrzeni ukrytej, zapewniając, że jest zgodna z określonym rozkładem.

A diagram of a computer code

Description automatically generatedRysunek 2 - Schemat działania VAE

Źrodło: Variational autoencoder. <https://en.wikipedia.org/wiki/Variational_autoencoder>.

### **Transformer Models**

Transformer Models, takie jak GPT-3, są stosowane do generowania tekstu, który jest nieodróżnialny od tego napisanego przez człowieka. Modele te są trenowane na dużych zestawach danych tekstowych i mogą generować długie, spójne fragmenty tekstu na podstawie krótkiego promptu.(7)

Modele transformacyjne są obecnie najnowocześniejszymi modelami NLP i są uważane za ewolucję architektury kodera-dekodera. Jednak, podczas gdy architektura kodera-dekodera polega głównie na sieciach neuronowych rekurencyjnych (RNNs) do ekstrakcji informacji sekwencyjnej, Transformery całkowicie pozbawione są tej rekurencji.

Są one specjalnie zaprojektowane do zrozumienia kontekstu i znaczenia poprzez analizę relacji między różnymi elementami, i prawie całkowicie polegają na technice matematycznej zwanej uwagą.

Pierwotnie wprowadzone w 2017 roku przez Google, modele transformacyjne są jednym z najnowszych i najbardziej wpływowych rozwojów w dziedzinie uczenia maszynowego. Pierwszy model Transformer został wyjaśniony w wpływowym artykule "Attention is All You Need".

Modele transformacyjne są wykorzystywane w wielu zastosowaniach, od tłumaczenia języka maszynowego, przez chatboty konwersacyjne, aż po zasilanie lepszych wyszukiwarek. Są one szczególnie skuteczne w rozwiązywaniu problemów sekwencyjnych, takich jak tłumaczenie maszynowe, generowanie tekstu i analiza sentymentu.

A diagram of a software algorithm

Description automatically generated

Rysunek 3 – Architektura Transformer Models

Żródło: The Transformer Model. https://machinelearningmastery.com/the-transformer-model/

## **Zastosowania Generatywnego AI**

Generatywne AI, mimo że jest stosunkowo nowym obszarem badań, już znalazło wiele zastosowań w różnych dziedzinach. W tym rozdziale uwaga będzie skupiona na zastosowaniach GenAI w kontekście cyberbezpieczeństwa i zarządzania ryzykiem.

GenAI ma potencjał do przekształcenia wielu aspektów cyberbezpieczeństwa, od wykrywania zagrożeń po reagowanie na incydenty. Może pomóc w automatyzacji wielu zadań związanych z bezpieczeństwem, co pozwala specjalistom ds. bezpieczeństwa skupić się na bardziej skomplikowanych problemach.

Jednak, jak każda potężna technologia, GenAI niesie ze sobą również swoje wyzwania i ryzyka. Może być wykorzystywane do celów złośliwych, takich jak tworzenie deepfakes, które mogą być używane do manipulacji i oszustw. Ponadto, ponieważ GenAI jest trenowane na dużych zestawach danych, istnieje ryzyko naruszenia prywatności, jeśli dane te zawierają informacje osobiste.

Rozdział ten pozwoli na przyjrzenie się bliżej różnym zastosowaniom GenAI w cyberbezpieczeństwie i zarządzaniu ryzykiem, a także potencjalnym wyzwaniom i ryzyku, które one niosą. Celem jest zrozumienie, jak GenAI może pomóc w poprawie cyberbezpieczeństwa i zarządzania ryzykiem, a także jakie są potencjalne zagrożenia i jak można je zminimalizować.

### **Wykrywanie i tworzenie ataków phishingowych**

Ataki phishingowe są jednym z najczęstszych rodzajów cyberprzestępstw. Polegają na wykorzystaniu oszukańczych e-maili i stron internetowych do kradzieży wrażliwych informacji, takich jak dane logowania czy informacje bankowe. (8)(9)

Generatywne AI może pomóc w walce z atakami phishingowymi na kilka sposobów. Oto przykładowe rozwiązania:

* **Automatyczne wykrywanie prób phishingu**: Generatywne AI może być używane do automatycznego wykrywania prób phishingu. Analizuje e-maile i strony internetowe pod kątem typowych cech ataków phishingowych. Może to obejmować analizę struktury e-maila, używanego języka, a nawet obrazów i logo.
* **Symulowanie ataków phishingowych**: GenAI może być używane do symulowania ataków phishingowych w celu testowania i poprawy środków bezpieczeństwa organizacji. Na przykład, może generować wiarygodnie wyglądające e-maile phishingowe, które mogą być używane do testowania, czy pracownicy są w stanie prawidłowo zidentyfikować takie próby.

Przykładem ataku phishingowego jest incydent, który dotknął Colonial Pipeline w 2021 roku. W wyniku ataku ransomware, który rozpoczął się od skutecznego ataku phishingowego, firma została zmuszona do zapłacenia milionów dolarów, aby przywrócić swoje systemy do działania.

Intruzi przeniknęli do sieci Colonial Pipeline, wykorzystując odsłonięte hasło do konta VPN - stwierdził Charles Carmakal, główny wiceprezes i CTO w firmie specjalizującej się w cyberbezpieczeństwie, Mandiant, podczas przesłuchania przed Komisją Bezpieczeństwa Krajowego Izby Reprezentantów 8 czerwca.

Spore grono firm stosuje VPN, aby umożliwić bezpieczny, zaszyfrowany dostęp zdalny do sieci firmowej. Zgodnie z zeznaniami Carmakala, jeden z pracowników Colonial Pipeline - którego tożsamość nie została ujawniona podczas przesłuchania - najprawdopodobniej użył identycznego hasła do VPN w innym miejscu. To hasło zostało w jakiś sposób naruszone podczas innego incydentu związanego z naruszeniem bezpieczeństwa danych.

Wiele rzeczy mogło zapobiec temu ataku. Na przykład, szkolenie pracowników z zakresu rozpoznawania podejrzanych e-maili mogłoby pomóc im uniknąć wpadki na wiadomości phishingowe. Ponadto, implementacja silnych polityk bezpieczeństwa, takich jak regularne aktualizacje oprogramowania, ograniczanie dostępu do wrażliwych systemów i danych tylko do niezbędnych użytkowników, oraz stosowanie wieloskładnikowego uwierzytelniania, mogłyby również pomóc w ochronie przed tego typu atakami.

Ważne jest, aby pamiętać, że ataki phishingowe są stale ewoluującym zagrożeniem, a cyberprzestępcy ciągle opracowują nowe techniki i taktyki, aby omijać środki bezpieczeństwa. Dlatego kluczowe jest utrzymanie aktualizacji zabezpieczeń, regularne szkolenie pracowników i utrzymanie świadomości zagrożeń cyberbezpieczeństwa na najwyższym poziomie.

Jednak, podobnie jak wiele innych technologii, GenAI ma również swoje ciemne strony. Może być wykorzystywane przez cyberprzestępców do tworzenia skomplikowanych ataków phishingowych, które są trudne do wykrycia dla tradycyjnych systemów bezpieczeństwa. Dlatego ważne jest, aby organizacje były świadome tych możliwości i podjęły odpowiednie środki, aby się przed nimi chronić.

### **Maskowanie danych i ochrona prywatności**

Maskowanie danych, znane również jako anonimizacja danych, jest techniką zabezpieczania, która polega na zastępowaniu wrażliwych danych danymi fikcyjnymi, ale realistycznymi. Celem maskowania danych jest ochrona prywatności i bezpieczeństwa informacji, jednocześnie umożliwiając korzystanie z danych w sposób, który nie narusza ich poufności.(10)(11)

Przykładowe działania, które można podjąć w celu implementacji maskowania danych, obejmują:

* **Identyfikacja wrażliwych danych**: Przed rozpoczęciem procesu maskowania danych, organizacje muszą zidentyfikować, które dane są wrażliwe i wymagają ochrony.
* **Wybór techniki maskowania danych**: Istnieje wiele technik maskowania danych, takich jak shuffling, substitution, encryption, i inne. Organizacje muszą wybrać technikę, która najlepiej pasuje do ich potrzeb.
* **Zastosowanie maskowania danych**: Po zidentyfikowaniu wrażliwych danych i wybraniu odpowiedniej techniki, organizacje muszą zastosować maskowanie danych, zastępując wrażliwe dane danymi fikcyjnymi.

W 2014 roku eBay doświadczył jednego z największych naruszeń danych w historii, kiedy hakerzy zaatakowali sieć firmy i skradli około 145 milionów rekordów użytkowników. Naruszenie obejmowało hasła, które były zaszyfrowane, a także adresy e-mail, daty urodzenia, adresy pocztowe i inne dane osobowe, ale nie dane finansowe, takie jak numery kart kredytowych.

eBay zalecił wszystkim użytkownikom natychmiastową zmianę haseł, ale eksperci ostrzegali, że nawet po zmianie haseł, naruszenie mogło mieć “katastrofalne” konsekwencje. Avivah Litan, analityk w firmie badawczej Gartner, powiedziała, że jeśli cyberprzestępcy zdołają skompilować dane z różnych źródeł, "w rurociągu jest potężny incydent, taki jak powszechne kradzieże tożsamości lub przejęcie tysięcy kont finansowych".

W odpowiedzi na atak, eBay zmusił wszystkich swoich użytkowników do zmiany haseł. Jednak eksperci ostrzegali, że to może nie być wystarczające, a konsekwencje ataku mogą być "katastrofalne". Ponad 15 milionów Brytyjczyków i ponad sto milionów osób na całym świecie było narażonych na kradzież tożsamości po ataku.

Gdyby eBay zastosowało techniki maskowania danych, mogłoby to zapobiec ujawnieniu wrażliwych informacji klientów, nawet w przypadku naruszenia bezpieczeństwa. Na przykład, mogliby zastosować technikę maskowania danych, aby stworzyć zestaw danych testowych, które są strukturalnie identyczne z prawdziwymi danymi, ale nie zawierają żadnych wrażliwych informacji.

### **Odpowiedź na incydenty**

Odpowiedź na incydenty, znana również jako reakcja na incydenty, odnosi się do działań, które organizacja podejmuje, gdy wierzy, że jej systemy informatyczne lub dane mogły zostać naruszone. Na przykład, profesjonaliści ds. bezpieczeństwa będą działać, jeśli zobaczą dowody na nieautoryzowanego użytkownika, złośliwe oprogramowanie lub awarię środków bezpieczeństwa.(12)(13)(14)

Celem odpowiedzi na incydenty jest jak najszybsze wyeliminowanie ataku cybernetycznego, odzyskanie, powiadomienie klientów lub agencji rządowych, zgodnie z wymogami prawnymi regionalnymi, oraz nauczenie się, jak zredukować ryzyko podobnego naruszenia w przyszłości.

Przykładowe działania, które można podjąć w celu implementacji odpowiedzi na incydenty, obejmują:

* **Utworzenie zespołu ds. odpowiedzi na incydenty (IRT)**: Zespół IRT składa się z osób odpowiedzialnych za reagowanie na incydenty bezpieczeństwa. Członkowie zespołu powinni mieć odpowiednie umiejętności i doświadczenie, a także dostęp do odpowiednich zasobów.
* **Utworzenie planu odpowiedzi na incydenty (IRP**): IRP to dokument, który opisuje, jak organizacja powinna reagować na incydenty bezpieczeństwa. Powinien zawierać szczegółowe procedury, które należy podjąć w odpowiedzi na różne typy incydentów.
* **Szkolenie i świadomość**: Wszyscy członkowie zespołu IRT, a także inni kluczowi pracownicy, powinni być regularnie szkoleni w zakresie odpowiedzi na incydenty. Szkolenia te powinny obejmować zarówno teoretyczne aspekty odpowiedzi na incydenty, jak i praktyczne ćwiczenia.

W 2014 roku Sony Pictures doświadczyło jednego z największych naruszeń danych w historii, kiedy grupa hakerów o nazwie “Guardians of Peace” włamała się do sieci Sony Pictures i ukradła ogromne ilości danych. Naruszenie obejmowało hasła, które były zaszyfrowane, a także adresy e-mail, daty urodzenia, adresy pocztowe i inne dane osobowe, ale nie dane finansowe, takie jak numery kart kredytowych.

Hakerzy użyli typu ataku znanego jako “wiper”, który niszczy dane na serwerach i instaluje ransomware. W wyniku ataku, ujawnione zostały poufne wewnętrzne e-maile, nieopublikowane filmy i wrażliwe dane pracowników, powodując szacowane straty na poziomie 15 milionów dolarów w bezpośrednim następstwie.

W odpowiedzi na atak, Sony Pictures podjęło szereg działań, w tym zamknięcie całej sieci komputerowej, aby zapobiec dalszemu rozprzestrzenianiu się ataku, a także współpracę z organami ścigania i firmami zewnętrznymi specjalizującymi się w cyberbezpieczeństwie, aby zidentyfikować sprawców i ocenić pełny zakres naruszenia.

Jednym z kluczowych elementów odpowiedzi na incydent jest szybkość reakcji. Im szybciej organizacja jest w stanie zidentyfikować i zareagować na incydent, tym mniejsze są potencjalne szkody. W przypadku Sony Pictures, firma była w stanie szybko zidentyfikować zagrożenie, izolować zainfekowane systemy i usunąć zagrożenie.

# **Ryzyka w Generatywnym AI**

## **Przegląd krajobrazu ryzyka Generatywnego AI**

Generatywne AI (Generative AI) to nowoczesna technologia, która rewolucjonizuje wiele dziedzin życia poprzez tworzenie nowych danych na podstawie istniejących wzorców za pomocą zaawansowanych algorytmów uczenia maszynowego. Technologia ta, obejmująca między wcześniej opisane enerative Adversarial Networks (GAN) oraz Variational Autoencoders (VAE), ma szerokie zastosowanie w różnych sektorach, takich jak sztuka, medycyna, rozrywka, edukacja i wiele innych. Mimo że generatywne AI przynosi wiele korzyści, wiąże się również z licznymi ryzykami, które należy dokładnie zrozumieć i zarządzać nimi w odpowiedni sposób.

Ryzyka związane z Generatywnym AI można podzielić na kilka głównych kategorii: techniczne, etyczne, prawne oraz społeczne. Każda z tych kategorii niesie ze sobą specyficzne wyzwania i zagrożenia, które mogą mieć znaczący wpływ na różne aspekty życia społecznego i gospodarczego.

* **Techniczne ryzyka** odnoszą się do samego funkcjonowania systemów generatywnych AI. Modele te mogą generować nieadekwatne lub błędne wyniki, co może prowadzić do nieprzewidywalnych konsekwencji. Ponadto, są one podatne na różne rodzaje ataków, takie jak ataki zakłócające (adversarial attacks), które mogą manipulować wynikami generowanymi przez AI. Awarie systemów, szczególnie w kontekstach krytycznych, jak medycyna czy transport, mogą prowadzić do poważnych konsekwencji.
* **Etyczne ryzyka** związane z generatywnym AI obejmują potencjalne nadużycia technologii do tworzenia szkodliwych lub nieetycznych treści. Przykładem mogą być deepfake'i, które mogą być wykorzystywane do manipulacji opinią publiczną, szantażu lub dezinformacji. Inne ryzyka etyczne obejmują generowanie mowy nienawiści, pornografii oraz innych nieodpowiednich treści, które mogą negatywnie wpływać na społeczeństwo. Te kwestie wymagają odpowiedniej regulacji i monitorowania, aby zapobiec ich szkodliwym skutkom.
* Generatywne AI wiąże się również z wieloma wyzwaniami prawnymi, w tym kwestiami własności intelektualnej i prywatności. Treści generowane przez AI mogą naruszać prawa autorskie, co prowadzi do pytań o to, kto jest właścicielem tych treści i jak je chronić. Ponadto, przetwarzanie danych osobowych przez AI bez odpowiedniej zgody stanowi naruszenie prywatności, co może prowadzić do problemów prawnych. Brak jasnych regulacji w tym obszarze może prowadzić do niepewności i sporów prawnych.
* **Społeczne ryzyka** generatywnego AI obejmują wpływ na dezinformację, zatrudnienie oraz zaufanie do informacji. Technologia ta może być używana do masowego tworzenia fałszywych informacji, co może prowadzić do destabilizacji społecznej i politycznej. Automatyzacja wielu zadań przez AI może skutkować utratą miejsc pracy w niektórych sektorach, co może wywoływać napięcia społeczne i ekonomiczne. Nadmierne poleganie na AI może również prowadzić do utraty zaufania do tradycyjnych źródeł informacji, co może mieć długofalowe konsekwencje dla społeczeństwa.

Podsumowując, Generatywne AI, mimo ogromnego potencjału, niesie ze sobą szereg ryzyk, które należy starannie analizować i zarządzać nimi w odpowiedni sposób. W kolejnych sekcjach szczegółowo przeanalizujemy poszczególne rodzaje ryzyk oraz ich potencjalne implikacje, aby lepiej zrozumieć, jak minimalizować negatywne skutki i maksymalizować korzyści płynące z tej innowacyjnej technologii.

## **Szczegółowa analiza potencjalnych ryzyk i ich implikacji**

### **Techniczne ryzyka**

Generatywne AI (Generative AI) niesie ze sobą szereg technicznych ryzyk, które wynikają z samej natury algorytmów i systemów, które je implementują. Poniżej omówimy szczegółowo najważniejsze z tych ryzyk oraz ich potencjalne implikacje.

1. Błędy i nieścisłości modeli

Modele generatywne, takie jak GAN (Generative Adversarial Networks) i VAE (Variational Autoencoders), mogą generować wyniki, które są błędne lub nieadekwatne. Te problemy mogą wynikać z kilku czynników:

* **Niedoskonałości danych treningowych**: Modele uczą się na podstawie dostępnych danych treningowych. Jeśli te dane są stronnicze, niepełne lub zawierają błędy, model będzie generował podobne błędy. Przykładowo, model trenujący na zdjęciach medycznych z jednego regionu świata może nie działać poprawnie na zdjęciach z innych regionów, prowadząc do błędnych diagnoz.
* **Błędy w algorytmach**: Nawet małe błędy w implementacji algorytmu mogą prowadzić do nieprzewidywalnych rezultatów. Modele mogą także źle interpretować dane wejściowe, generując nieadekwatne wyjścia.
* **Nieprzewidywalne interakcje między komponentami systemu**: Złożone systemy AI składają się z wielu współpracujących komponentów. Problemy mogą pojawić się, gdy te komponenty nie współpracują ze sobą zgodnie z oczekiwaniami, prowadząc do błędów systemowych.

Przykład: W kontekście medycznym, model generujący obrazy MRI może błędnie interpretować artefakty na obrazach jako patologie, prowadząc do błędnych diagnoz i potencjalnie szkodliwego leczenia.

1. Ataki zakłócające (Adversarial Attacks)

Modele generatywne są podatne na ataki zakłócające, które polegają na wprowadzaniu małych, precyzyjnie zaprojektowanych zakłóceń do danych wejściowych w celu spowodowania błędnych wyników modelu.

* **Techniki ataków zakłócających**: Te ataki mogą obejmować dodanie niewielkiego szumu do obrazów, który jest niezauważalny dla ludzkiego oka, ale powoduje błędne klasyfikacje przez model AI.
* **Potencjalne konsekwencje**: Ataki zakłócające mogą być wykorzystywane do manipulacji systemami rozpoznawania twarzy, autonomicznymi pojazdami czy systemami bezpieczeństwa. Na przykład, zakłócenie obrazu drogowego może sprawić, że autonomiczny pojazd zinterpretuje znak stopu jako znak ograniczenia prędkości, co może prowadzić do wypadku.

Przykład: W przypadku systemów rozpoznawania twarzy, niewielkie zmiany w obrazie mogą spowodować, że osoba nie zostanie rozpoznana lub zostanie błędnie zidentyfikowana jako ktoś inny, co ma poważne implikacje dla bezpieczeństwa.

1. Awarie systemów

Złożoność systemów AI może prowadzić do trudności w ich utrzymaniu i zarządzaniu, co zwiększa ryzyko awarii. Awarie te mogą mieć różne przyczyny:

* **Błędy w oprogramowaniu**: Złożone systemy AI składają się z wielu komponentów programistycznych. Nawet małe błędy w jednym z tych komponentów mogą prowadzić do awarii całego systemu.
* **Złożoność operacyjna**: Zarządzanie i utrzymanie złożonych systemów AI wymaga zaawansowanej wiedzy i umiejętności. Brak odpowiedniego zarządzania może prowadzić do nieprzewidywalnych awarii.
* **Awaria infrastruktury**: Systemy AI często opierają się na rozproszonej infrastrukturze chmurowej. Problemy z infrastrukturą, takie jak awarie serwerów czy problemy z łącznością, mogą prowadzić do przerw w działaniu systemów AI.

Przykład: W systemach autonomicznych pojazdów, awaria systemu może prowadzić do utraty kontroli nad pojazdem, co stanowi bezpośrednie zagrożenie dla życia pasażerów i innych uczestników ruchu drogowego.

# **Bezpieczeństwo w Generatywnym AI.**

# **Rozdział 4.**

# **Rozdział 5.**

# **Bibliografia**

|  |  |
| --- | --- |
| [1] | ValueLabs, „Empowering Cybersecurity with Generative AI,” 2023. [Online]. Available: https://www.valuelabs.com/resources/blog/cybersecurity/cybersecurity-and-generative-ai/. |
| [2] | geeksforgeeks, „Variational AutoEncoders,” 2023. [Online]. Available: https://www.geeksforgeeks.org/variational-autoencoders/. |
| [3] | J. Rocca, „Understanding Variational Autoencoders (VAEs),” 2019. [Online]. Available: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73. |
| [4] | J. Schulze, „What is GAN? Generative Adversarial Networks Explained,” 2024. [Online]. Available: https://www.coursera.org/articles/what-is-gan. |
| [5] | G. v. d. Linden, „THE TRANSFORMATIVE POWER OF GENERATIVE AI IN CYBERSECURITY,” 2024. [Online]. Available: https://www.capgemini.com/insights/expert-perspectives/the-transformative-power-of-generative-ai-in-cybersecurity/. |
| [6] | J. Ferrer, „How Transformers Work: A Detailed Exploration of Transformer Architecture,” 2024. [Online]. Available: https://www.datacamp.com/tutorial/how-transformers-work. |
| [7] | Z. Alkhalil, I. Khan i L. Nawaf, „Phishing Attacks: A Recent Comprehensive Study and a New Anatomy,” 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2021.563060/full. |
| [8] | S. M. Kerner, „Colonial Pipeline hack explained: Everything you need to know,” 2022. [Online]. Available: https://www.techtarget.com/whatis/feature/Colonial-Pipeline-hack-explained-Everything-you-need-to-know. |
| [9] | The Week, „eBay hack 'one of the biggest data breaches in history',” 2014. [Online]. Available: https://theweek.com/technology/58624/ebay-hack-one-of-the-biggest-data-breaches-in-history. |
| [10] | IBM, „What is incident response?,” 2024. [Online]. Available: https://www.ibm.com/topics/incident-response. |
| [11] | Microsoft, „What is incident response?,” 2024. [Online]. Available: https://www.microsoft.com/en-us/security/business/security-101/what-is-incident-response. |
| [12] | L. Pryimenko, „7 Examples of Real-Life Data Breaches Caused by Insider Threats,” 2024. [Online]. Available: https://www.ekransystem.com/en/blog/real-life-examples-insider-threat-caused-breaches. |
| [13] | A. Richman, „5 Real-World Data Masking Examples,” 2023. [Online]. Available: https://www.k2view.com/blog/data-masking-examples. |
| [14] | W. J. B. M. G. T. L. M. Yagmur Yigit, „Review of Generative AI Methods in Cybersecurity,” 2024. [Online]. Available: https://arxiv.org/html/2403.08701v1. |