ale także w atakach na systemy analizy sentymentu czy podczas oceny ryzyka kredytowego. Fikcyjne osobowości mogą zostać uwiarygodnione przez generowanie fikcyjnych działań czy przez publikowanie zdjęć, zawierających nieistniejące osoby (rys. 3.7). Podejście to, zwłaszcza połączone z atakiem typu *shilling*, jest szczególnie trudne do wykrycia (Bhaumik i in., 2006).

Rysunek 3.7. Sztucznie wygenerowane, fikcyjne fotografie ludzi, przy zastosowaniu techniki GAN. Utworzone w ten sposób dane są praktycznie nieodróżnialne od rzeczywistych

Źródło: https://thispersondoesnotexist.com/ (dostęp: 27.05.2020 r.).

3.4.3.5. Inne zagrożenia

Dotychczasowe klasyfikacje zagrożeń wynikających z antagonistycznego uczenia maszynowego opierają się głównie na dwóch kategoriach: na czasie, w którym atak został wykonany (infekcyjny, inwazyjny lub atak na klasyfikator), lub na poziomie wiedzy dostępnej dla atakującego (*black box* lub *white box*). Można także dokonać klasyfikacji wybranych zagrożeń na podstawie procesów biznesowych, będących celem, ataku lub według stosowanych w nich technikach AI (tab. 3.2).

Tabela 3.2. Zagrożenia wynikające z antagonistycznego uczenia maszynowego

Biznesowe zastosowanie AI	Przykłady zagrożeń
Identyfikacja nadużyć	Manipulowanie danymi, aby ukryć nielegalną działalność, związaną przykładowo z nadużyciami finansowymi lub praniem brudnych pieniędzy. Generowanie próbek antagonistycznych służy w tym przypadku dwóm celom: zastąpieniu podejrzanej transakcji inną transakcją (wygenerowaną sztucznie) lub obudowaniu nadużycia innymi transakcjami (także sztucznymi), aby nadużycie nie było traktowane jak anomalia (Schreyer i in., 2019).
Bezpieczeństwo danych	Ukrywanie faktu kradzieży danych z systemów informatycznych. Systemy identyfikacji nadużyć wykrywają działania pracowników, które odbiegają od normy (np. uruchamianie kilkadziesiąt razy tego samego raportu, zawierającego dane klientów, podczas gdy inni pracownicy uruchamiają go średnio raz w tygodniu). Atak polega na przygotowaniu robota programowego, aby wykonywał on działania symulujące pracownika, jednak prowadzące do pozyskania jak największej ilości danych.
Zarządzanie portfelem inwestycyjnym	Wprowadzenie w błąd systemów realizujących automatyczne transakcje finansowe, przez wykorzystanie luk w regułach działania tych systemów. Generowanie dużej liczby transakcji powodujące, że systemy zaczynają je intepretować według zaimplementowanych reguł, co może prowadzić do zmian w kursach akcji lub walut. Przykładowo, w 2015 r. rosyjscy hakerzy dokonali ataku na sektor finansowy, wykorzystując tę właściwość robotów. Hakerzy wykorzystali złośliwe oprogramowanie, aby na krótko zdestabilizować kurs wymiany rubla do dolara (Hacker News, 2016).
Symulacje finansowe	Wprowadzenie fałszywych danych transakcyjnych do uczącego zbioru danych, aby wprowadzić w błąd systemy symulacyjne. Atakujący może w ten sposób wpłynąć na parametry opracowanego modelu symulacyjnego. Modele te są regularnie szkolone, aby uwzględnić nowsze dane, co czyni je podatnymi na tego typu ataki (Cantos, 2019).
Zarządzanie ryzykiem kredytowym	Wprowadzenie w błąd systemu oceny ryzyka kredytowego, przez prezentowanie spreparowanych lub zmodyfikowanych danych. Taki system może błędnie oszacować ryzyko kredytowe i sprawić, że bank podejmie niepożądane działania i np. udzieli kredytu podmiotowi niewypłacalnemu.

Źródło: opracowanie własne.