Česká asociace
umělé inteligence

# DEEPFAKE 2024

## OBRANNÁ STRATEGIE PRO ČESKÉ FIRMY

# Předmluva

Vážení,

celosvětový počet bezpečnostních incidentů s deepfake napříč odvětvími vzrostl v roce 2023 na desetinásobek ve srovnání s rokem předchozím. Jen v Evropě se jedná o **780% nárůst** počtu detekovaných případů. Není to náhoda. Rychlý rozvoj umělé inteligence, kterého jsme byli v roce 2023 svědkem, mění pravidla hry. Hrozby spojené se sofistikovanými útoky a syntetickými médii budou v roce 2024 patřit k jedněm z nejpalčivějších problémů v kontextu bezpečnostní politiky firem.

Cílem manuálu **DEEPFAKE 2024: OBRANNÁ STRATEGIE PRO ČESKÉ FIRMY** je poskytnout zaměstnancům českých firem a organizací srozumitelný přehled o současných i nastupujících trendech v oblasti deepfakes, o možnostech jejich detekce a autentizace, a o strategiích, které mohou využít k obraně proti těmto hrozbám. Je nezbytné, aby si české firmy uvědomily, že deepfakes nejsou jen problémem velkých korporací nebo politických organizací.

**Česká asociace umělé inteligence** se tímto manuálem snaží poukázat na důležitost vzdělávání zaměstnanců a budování povědomí o těchto hrozbách ve firmách. Pevně věřím, že dokument poskytne **cenné informace** pro všechny, kteří se chtějí aktivně podílet na ochraně svých organizací v digitálním světě roku 2024 a dále.

S pozdravem

**Lukáš Benzl**

ředitel České asociace umělé inteligence

# Poděkování

Tento manuál by nevznikl bez podpory **našich hlavních partnerů** a členů asociace.

**Fraud management** našeho partnera **Analytics Data Factory** přináší služby v oblasti systémů na detekování fraudů. Zahrnuje nejen kompletní analýzu nastavení procesů, ale i následnou podporu jejich řízení, a to včetně kompletní správy a rozvoje technologií, které výrazně usnadňují **odhalování nestandardního chování**. Děkujeme za pomoc s přípravou obsahu tohoto manuálu.

**EY Česká republika** nabízí širokou škálu služeb včetně poradenství v oblasti inovací, digitálních řešení, **IT a rizik**, finančního, právního a daňového poradenství nebo poradenství v oblasti fúzí a akvizic.

**Advokátní kancelář Legitas**, spoluzakladatel České asociace umělé inteligence, se specializuje na moderní trendy v technologiích **včetně umělé inteligence** a e-commerce. V případě zájmu si neváhejte **sjednat konzultaci**.

# Obsah

# 1. Úvod do světa deepfakes

Hrozby spojené s deepfakes představují stále rostoucí výzvu pro všechny firmy využívající moderní technologie. Rostoucí dostupnost aplikací umělé inteligence a efektivita technik syntetických médií vede k častějšími a sofistikovanějším útokům. Deepfakes jsou **zvláště znepokojivým typem syntetických médií**, která využívají umělou inteligenci k vytváření uvěřitelných a vysoce realistických médií.

Největší hrozby zneužití zahrnují techniky, které ohrožují značku organizace, **napodobují vedoucí pracovníky a používají podvodnou komunikaci** k získání přístupu k sítím, komunikaci a citlivým informacím organizace.

Organizace mohou podniknout řadu kroků k identifikaci a obraně proti deepfakes. V roce 2024 je na místě zvážit implementaci technologií pro detekci deepfakes a určení původu médií včetně schopností ověřování v reálném čase, pasivních detekčních technik a ochrany prioritních zaměstnanců a jejich komunikace.

I vaše firma může podniknout **relativně jednoduché aktivity** k minimalizaci dopadu škodlivých technik deepfake včetně nacvičování reakcí na pokusy o zneužití. Phishing, tedy typ kybernetického útoku zneužívající **techniky sociálního inženýrství**, bude s rozvojem AI ještě větší výzvou než kdykoliv dřív.

Sociální inženýrství je metoda manipulace lidí za účelem získání důvěrných informací nebo k provedení určitých akcí, které jsou výhodné pro útočníka. Tato technika **často zahrnuje psychologické triky**, jako je vytváření důvěry, využívání autority, vzbuzování strachu nebo naléhání na okamžitou reakci. Cílem je přimět jednotlivce k neúmyslnému prozrazení citlivých dat, jako jsou hesla, finanční informace nebo přístupové kódy, nebo k provádění akcí, které by jinak neprovedl, jako je otevření škodlivého odkazu nebo převod peněz. Sociální inženýrství je často využíváno v kybernetických útocích a podvodech.

Nástroje a techniky, které lze použít k manipulaci s autentickými multimédii, existují již desetiletí. Nicméně rozsah používání manipulace s médii v roce 2023 **dramaticky vzrostl**, protože složitost manipulace médií klesla na minimum.

Vytvoření sofistikovaného falza pomocí specializovaného softwaru dříve mohlo profesionálovi trvat od několika dní do několika týdnů, ale dnes lze velmi kvalitní deepfake vytvořit na chytrém telefonu **za zlomek času** s omezenými nebo žádnými technickými znalostmi.

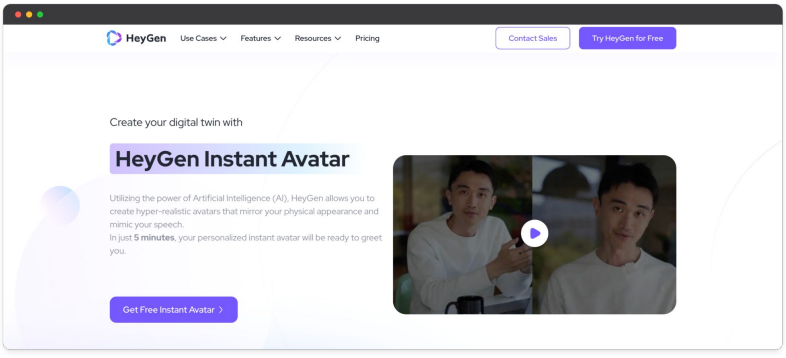
„Viníkem“ je pokrok v oblasti hlubokého učení. Ten nejen usnadňuje tvorbu falešných multimédií, ale také činí jejich hromadnou výrobu méně nákladnou. Generativní AI jako velké jazykové modely a difuzní modely (nebo jejich kombinace) umožňuje vytvářet **vysoce realistická syntetická multimédia** na základě mnohem větších datových sad.

Demokratizace AI nástrojů pro tvorbu deepfakes se dostala na seznam největších rizik pro rok 2024. Trh je zaplaven volně dostupnými nástroji, které činí tvorbu nebo manipulaci s multimédii **otázkou několika minut**. To ostatně potvrzuje i člověk, který jako jeden z prvních v ČR ukázal, jak snadné je používat aplikace jako **HeyGen**.

„AI deepfake je pro mě poměrně zajímavé téma, i z toho důvodu, že jsem byl v Česku možná jedním z prvních lidí, kterým se podařilo úspěšně vytvořit a následně i zveřejnit vlastního AI avatara. Video jsem vytvořil během hodiny jen s minimálním vybavením a v podmínkách, které mají opravdu daleko k profesionálnímu studiu. Výsledkem byla moje velmi realistická video kopie. Ta může mluvit například čínsky, nebo jakýmkoliv jiným jazykem," přibližuje tvorbu AI avatara **Dan Gottwald**, produktový manager AI ve Vltava Labe Media a člen České asociace umělé inteligence.

„Je děsivé, jak snadno lze vytvořit svou kopii a následně ji použít k mnoha různým účelům. Momentálně také testuji klonování hlasu, které je téměř nerozeznatelné od originálu. Stačí mít 2 minuty nahrávky nějaké staženého hlasu a během pár vteřin uděláte kopii. Asi si dovedete představit jak (s trochou sociálního inženýrství) dokonalá je to zbraň pro různé podvodníky a vyděrače. Měli bychom si velmi pečlivě hlídat, kam dáváme svoje fotky či videa a kdo k nim má přístup," dodává Dan.

Demokratizace AI nástrojů pro tvorbu deepfakes se dostala na seznam **největších rizik pro rok 2024**. Čeští zaměstnanci se musí připravit na dobu, kdy budou intenzivně testováni falešnými online účty vydávajícími se za spolupracovníky, podvodnými hlasovými zprávami od „nadřízených" či falešnými video záběry. Nástroje jako zmiňovaný HeyGen mají sloužit k tvorbě marketingových či například onboardingových videí, ale velmi snadno díky nim vytvoříte také deepfake s někým úplně cizím, kdo o tom nemusí ani vědět.



# 2. Dokážete ihned odhalit deepfake?

Existuje několik termínů, které se používají k popisu médií, která byla synteticky generována nebo upravena. Mezi historicky nejběžnější patřily tzv. shallowfakes. Oproti deepfakes se **nejedná o sofistikovanou technologii**. Je to pouhá manipulace s obrazem prostřednictvím editace nebo záměrným vytržením z kontextu.

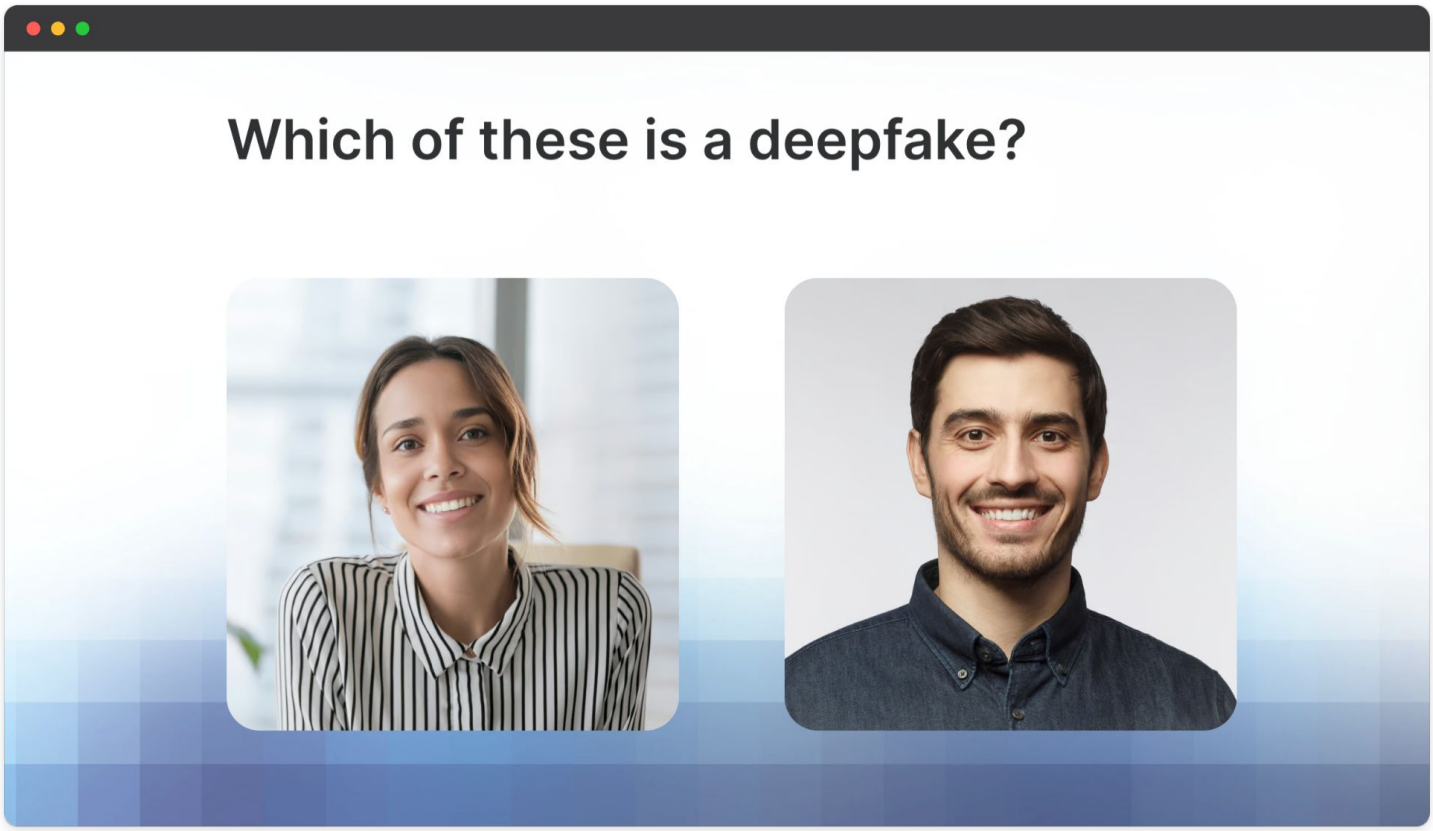
Pojďme si však přiznat, že i shallowfakes v mnoha případech mohou být stejně efektivní jako technicky sofistikovanější metody. Mezi konkrétní příklady patří například zpomalení videa přidáním opakujících se snímků, aby to vypadalo, jako by byl jedinec na videu intoxikován. Další metodou je kombinace zvukových klipů z jiného zdroje a nahrazení zvuku ve videu, aby se změnil příběh.

Na druhou stranu multimédia, která byla vytvořena (plně syntetická) nebo upravena (částečně syntetická) pomocí umělé inteligence, **se označují jako deepfakes**. Mezi **aktuální příklady**, na které jste mohli v médiích narazit, patří:

- Spoluzakladatel antivirové firmy Avast a jeden z nejbohatších Čechů Pavel Baudiš představuje investiční projekt Avast Capital v podvodné online reklamě.
- Snímek zobrazující výbuch poblíž Pentagonu, který byl sdílen v květnu 2023, způsobil obecné zmatky a turbulence na akciovém trhu.
- V období před vysoce kontroverzními parlamentními volbami v roce 2023 na Slovensku se po řetězových e-mailech šířilo **deepfake audio**, na kterém měl lídr strany Progresívne Slovensko Michal Šimečka diskutovat o plánech zmanipulovat volby.
- Britský občanský žurnalista Eliot Higgins vytvořil pomocí Midjourney obrázky Trumpa, jak se násilně brání zatčení, utíká před newyorskou policejní jednotkou a je ve vězení v oranžové kombinéze.

Podle průzkumu společnosti iProov **43 % globálních respondentů** připouští, že by nebyli schopni rozeznat rozdíl mezi skutečným videem a deepfake videem.

Společnost iProov stojí také za kvízem „**Can You Spot a Deepfake?**“, který si nyní můžete zdarma vyzkoušet. Kromě několik obecných otázek budete mít možnost vyzkoušet si, **zdali poznáte deepfake**. A věřte, že to není vůbec jednoduché.




Výzkumníci z University College London v roce 2023 použili algoritmus převodu textu na řeč trénovaný na dvou veřejně dostupných souborech dat, jeden v angličtině a druhý v mandarínštině, aby vygenerovali 50 vzorků deepfake řeči v každém jazyce. Ukázky zvuku byly přehrány 529 účastníkům. Účastníci byli schopni identifikovat falešnou řeč **pouze v 73 % případů**. Toto číslo se mírně zlepšilo poté, co účastníci absolvovali školení k rozpoznání aspektů deepfake řeči.

Říkáte si, že poznat deepfake vlastníma očima a ušima bude v roce 2024 prakticky nemožné ? Máte pravdu, ale **to neznamená**, že nebude možné se bránit.

„Základním pravidlem je ověřovat informace z více zdrojů a mít na paměti rizika prostředí internetu. Musíme mít představu, z jakých zdrojů můžeme očekávat AI generovaný obsah. Do budoucna se vyplatí začít používat AI rozšíření do webových prohlížečů. Ta budou schopna takový obsah detekovat a upozornit nás na něj," nastiňuje budoucí vývoj Jack Szabo, lead developer ve společnosti **Verisoft**.

Dobrou zprávou je, že vědci a inženýři si uvědomují potenciální škody, kterou mohou způsobit hyperrealistická syntetická média, a proto neúnavně pracují na inovativních řešeních, a to nejen v podobě výše zmíněných rozšíření do prohlížečů. Navrhované nástroje často využívají **pokročilé algoritmy strojového učení** samy o sobě a snaží se přechytračit a identifikovat deepfakes v neustále se vyvíjejícím prostředí syntetických médií.

Vyvracení deepfake útoků a rozpoznávání deepfakes se může zdát jednoduché a logické. Toto logické hledisko však ignoruje fakt, že se nejedná o spor mezi skutečností a fikcí, ale pravdou a uvěřitelností. Něco nemusí být pravda, pokud je to uvěřitelné. Lidé věří věcem, které opakovaně slyší a vidí. Výsledkem je útok, který může ohrozit vaši firmu a přetrvávat déle, než byste předpokládali.

Je nezbytné si uvědomit, že boj proti deepfakes není pouze technologický. Jak se vyvíjí technologie, mění se i strategie používané osobami se zlými úmysly. K doplnění vývoje sofistikovaných nástrojů jsou proto potřeba vzdělávací programy zaměstnanců. Veřejné pochopení existence a potenciálních nebezpečí deepfakes je v tomto boji mocnou zbraní. Vzdělávání umožňuje jednotlivcům **kriticky vyhodnotit informace**, se kterými se setkávají, a podporuje společnost méně náchylnou k manipulaci.

# 3. Typy častých deepfake útoků na firmy

Veřejné obavy ohledně syntetických médií zahrnují jejich využití v dezinformačních operacích, které jsou navrženy tak, aby ovlivnily veřejnost a šířily nepravdivé informace o politických, sociálních, vojenských nebo ekonomických otázkách a způsobily zmatek, nepokoje a nejistotu. Nicméně hrozby syntetických médií, se kterými se organizace nejčastěji setkávají, zahrnují aktivity, které mohou poškodit značku, zcizit peníze, nebo zásadně narušit bezpečnost samotné organizace. Pojďme se podívat na **typy častých deepfake útoků na firmy**, se kterými se zaměstnanci budou v roce 2024 setkávat nejčastěji.

## Napodobení hlasu nadřízeného v telefonu

Útočníci mohou používat deepfakes, které zahrnují manipulované audio, k pokusu o napodobení výkonných pracovníků organizace a dalšího vysoce postaveného personálu. Technologie deepfake **dokáže klonovat hlas**, což je běžné způsob ověřování a autorizace aktivity. Používání technologie deepfake ke klonování hlasu a vydávání se za jednotlivce má vést k podvodným finančním transakcím. Tato praktika je obzvláště nebezpečná **v kombinaci se spoofingem**.

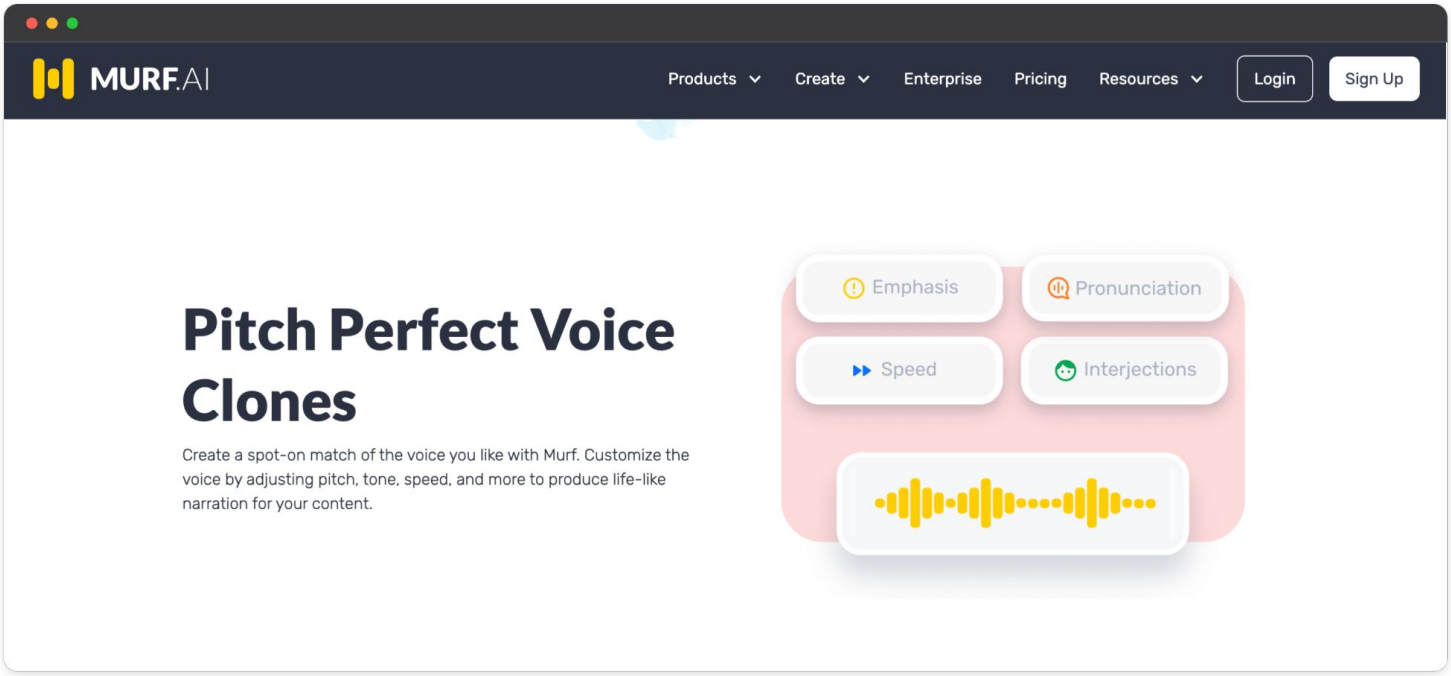
Při spoofingu se stane, že na displeji vašeho telefonu při příchozím hovoru uvidíte číslo, **které vypadá důvěryhodně**, ale ve skutečnosti za ním stojí někdo cizí, kdo pouze předstírá, že je někým jiným.

Útočníci jsou schopni imitovat libovolné telefonní číslo, takže vám mohou zavolat například pod záminkou, že volají z vaší banky. Nebo napodobí číslo vašeho nadřízeného. Některé firmy v České republice mají telefonní čísla zaměstnanců veřejně dostupná na webu. Zjistit číslo konkrétní vedoucí osoby tak není problém.

Pokud narazíte na spoofing, neváhejte kontaktovat svého telefonního operátora a informujte ho o podrobnostech podvodného hovoru. Operátoři v takových situacích **mohou příslušný provoz zablokovat**. Jen T-Mobile v České republice denně detekuje a zablokuje přes 4000 podezřelých aktivit.

- Zákazníci O2 mohou zavolat na infolinku 800 02 02 02
- Zákazníci Vodafonu mohou poslat informace o hovoru (číslo volajícího a čas události) bezplatnou SMSkou na číslo 7726
- Zákazníci operátora T-Mobile mohou volat na infolinku 800 73 73 73

V České republice je trh telekomunikací regulován a operátoři mají přístup k informacím o identitě volajících čísel. To znamená, že v případě podvodu je možné zjistit, kdo hovor inicioval. Tato možnost sledování však platí **jen pro hovory uskutečněné z České republiky**. Většina podvodných telefonátů je realizována ze zahraničí.



## Deepfakes v online schůzkách a pohovorech

Útočníci mohou použít stejné typy manipulovaných mediálních technik a technologií pro získání přístupu k personálu, operacím a informacím organizace. Tyto techniky mohou zahrnovat použití manipulovaných médií během **virtuálních pracovních pohovorů** (zejména pro vzdálené pozice) a online schůzek.

Úspěšné pokusy mohou aktérům umožnit získat citlivé finanční, majetkové nebo interní bezpečnostní informace. Techniky používané k **napodobení konkrétních zákazníků** mohou být také využívány k získání přístupu k jednotlivým zákaznickým účtům pro přístup k účtu nebo jiné účely shromažďování informací.

V srpnu 2023 zaměstnanci slovenské společnosti GymBeam přišla zpráva na WhatsAppu z účtu imitující účet Dalibora Cicmana, zakladatele a ředitele GymBeam. Ve zprávě bylo sdělení, že je nutné se rychle spojit, a odkaz na Microsoft Teams. Na videohovoru byla **deepfake kopie Cicmana s další osobou, která byla představena jako externí právník**. Během schůzky se falešná dvojice snažila zjistit zůstatky na firemních bankovních účtech. „Šťastnou náhodou bylo, že jsme se daný den osobně setkali a obratem jsem odepisoval na Slacku. Prý kopie vypadala velmi věrohodně a podezřelé bylo jen to, proč naléhavě potřebuji informace o bankovních účtech přes Teams, když využíváme Hangouts nebo Slack," popisuje Cicman.

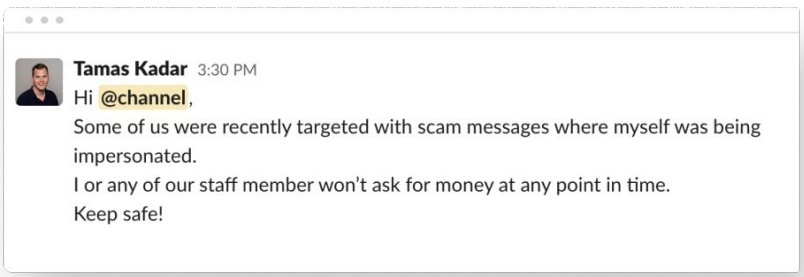
Deepfake útok může začít z jakékoliv komunikační platformy, včetně e-mailů nebo falešných profilů na sociálních sítích. Útočníci mají možnost vytvořit přesvědčivé podvodné komunikace, **které se zdají být od skutečných osob**. V rámci těchto útoků mohou být využívány i soukromé fotografie osoby, kterou se útočník snaží napodobit.

To znamená, že například v rámci videohovoru nebo na fotografii vytvořené pomocí deepfake technologie může být jako pozadí použita domácí pracovna dané osoby, což může u oběti útoku vyvolat falešný dojem skutečné komunikace. A ruku na srdce. Většina lidí včetně vedoucích pracovníků **si své soukromí nehlídá**. V období pandemie navíc většina z nás nějakým způsobem během online konverzace sdílela své domácí prostředí. Natočili jste někdy doma webinář a nahráli ho na YouTube? Útočník tento záznam může snadno zneužít.

## Zveřejnění falešného firemního prohlášení

Zaměstnanec pomlouvající zaměstnavatele, manažer obviňující partnery či finanční ředitel oznamující bankrot. Takový deepfake může **markantně poškodit reputaci a značku**. V době, kdy vaše firma na tuto hypotetickou událost zareaguje, bude pravděpodobně pozdě. Tento druh dezinformací o vaší značce bude na internetu viset roky. A nejen to.

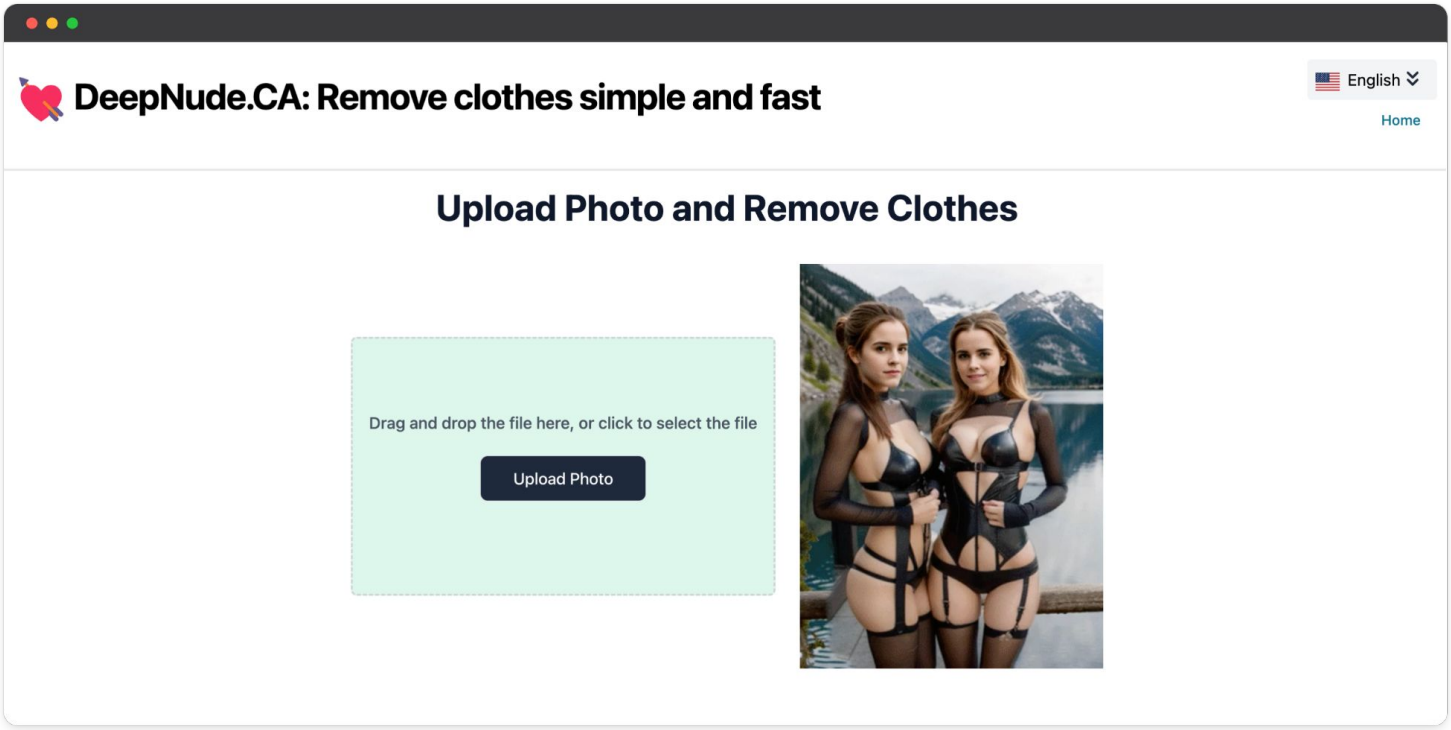
Ceny akcií kolísají na základě událostí. Například když kvalitní vedoucí pracovník odejde z veřejně obchodované společnosti, cena akcií se sníží kvůli obavám z nového vedení. **Deepfake generálního ředitele**, který oznámil odchod finančního ředitele, může způsobit, že akcie utrpí krátký pokles ceny. I když to může působit zanedbatelně, při správném načasování má takový dopad na odměny zaměstnanců či úsilí společnosti o získání investice.



## Vydírání zaměstnanců deepfake nahotou

Deepfake hrozba však nemusí přijít pouze z vnějšího prostředí. Historie a vznik deepfake technologií jsou neodmyslitelně spojeny s jejich využíváním pro tvorbu pornografie **bez souhlasu postižených osob**. Představte si situaci, kdy zaměstnanec vytvoří takovýto falešný obsah s využitím podoby jiného zaměstnance a rozšíří ho.

Tento scénář může znít jako hypotéza, ale FBI již informovala o případech, kdy byla použita podobná metoda cíleně proti jednotlivcům. Tento druh útoku je zákeřný a zlomyslný, **poškozuje psychické zdraví zaměstnanců,** narušuje jejich kariérní postup a téměř jistě vede k nákladným soudním sporům. Navíc se může zaměřit i na vyšší manažery, členy představenstva a další významné postavy.




Problematice deepnude se Česká asociace umělé inteligence podrobně věnovala v článku **Varování: Svlékací aplikace a AI deepnude! Jak reagovat v případě zneužití fotografie?**

# 4. Příklady konkrétních útoků a postupů

V roce 2019 utrpěla společnost Empresa Municipal de Transportes de Valencia podvod ve výši 4 milionů EUR. Jak se to stalo? Skupina kyberzločinců se uchýlila k oblíbené technice sociálního inženýrství posledních let, tzv. podvodu s generálním ředitelem. Prostřednictvím předstírání identity prostřednictvím e-mailu a provádění **falešných telefonních hovorů** se zločincům podařilo přimět provozního ředitele, aby vydal příkaz k provedení až osmi převodů v hodnotě 4 milionů EUR za účelem uskutečnění údajné akvizice v Číně.

Další případ se stal o rok později, v roce 2020, během koronavirové krize. Zendal Pharmaceuticals se stala obětí podvodu v hodnotě 9,7 milionu EUR. Operace zločinců byla podobná. Vydávali se za generálního ředitele společnosti, aby pověřil finančního manažera provést převody za účelem akvizice. Vydávali se také za profesionály z poradenské firmy. Deepfake s generálním ředitelem je prostě sofistikovanější a ambicióznější verze nejpopulárnější techniky sociálního inženýrství - **phishingu**. Základní mechanika tohoto typu útoku je následující.

1. Útočníci si nastaví cíl a nastudují si strukturu organizace.
2. Následně se vytvoří **falešná e-mailová adresa** tak, aby vypadala legitimně, například obsahuje doménu společnosti.
3. E-mail je zaslán manažerovi s pravomocí realizovat velké peněžní převody.
4. Když je podvod s generálním ředitelem na „dobré“ cestě, vygeneruje se dokumentace, jako jsou faktury nebo smlouvy, a pošle se oběti, aby podvod vypadal bezúhonně.
5. Když společnost nebo banky podvod konečně odhalí, **zločinci zmizí** a peníze se přesouvají do různých zemí, aby bylo obtížné je získat zpět.

# 5. Proč jsou deepfake útoky úspěšné?

AI deepfakes dramaticky **zvyšují šanci na úspěšný phishingový útok**, protože zaměstnanci mohou být snáze oklamáni a přiměni k poskytnutí citlivých informací nebo k provedení neautorizovaných finančních transakcí. Podvodníci vědí, že potřebují hromadně posílat phishingové zprávy, aby našli oběť. Dříve jim to trvalo hodiny nebo dny. Díky AI a velkým jazykovým modelům se tento čas **značně zkrátil**. Pomocí LLM mohou podvodníci vytvořit realistickou zprávu a poté kontaktovat uživatele ve velkém objemu.

Ačkoli má každá zločinecká skupina svou metodologii a taktiky, všechny útoky využívající podvody **kombinují 3 prvky**, které jim umožňují:

1. Obcházet velmi šikovně podezření oběti
2. Nutit k jednání bez pečlivosti, často nad rámec pravomocí
3. Zabránit oběti v interakci s ostatními ve firmě

Toto jsou **3 kritické předpoklady** pro úspěšný podvod:

1. Příkaz zadává zpravidla nadřízený, a proto mnoho lidí, i když zastávají manažerské funkce, ho nezpochybňuje
2. Příkaz apeluje na diskrétnost zaměstnance při řešení finančních transakcí jako akvizice či kontrast
3. Manipulace je doplněna klasickým prvkem útoků sociálního inženýrství, kterým je spěch a urgence

Z čistě psychologického úhlu pohledu je být součástí malého jádra lidí, kteří o transakci vědí, **zdrojem pocitu nezbytnosti**. Oběť díky tomuto pocitu ignoruje varovné signály. Připadá si důležitá.

# 6. Nadcházející deepfake trendy

Útoky využívající deepfake nadřízené představují v době rozvoje generativní AI jednu z největších hrozeb. Vývoj technologií pro tvorbu syntetických médií se ubírá směrem, který **snižuje náklady a odstraňuje technické překážky** pro jejich zneužití.

Očekává se, že do roku 2030 trh s generativní AI dosáhne hodnoty přes 100 miliard dolarů, s průměrným ročním růstem **přesahujícím 35 %**. Ačkoli schopnosti zneužívání těchto technologií neustále rostou, zdokonalují se i metody obrany, jako například identifikace a omezení deepfake obsahu. Nicméně detektory AI generovaného obsahu mohou mít **problémy s falešně pozitivními výsledky**, když označí lidský text za generovaný AI. Očekává se technologický závod mezi vývojem syntetických médií a schopnostmi detekce AI generovaného a ověřování pravého obsahu.

`Velikost trhu s deepfake softwarem byla v roce 2023 oceněna na částku 72,41 milionu USD a očekává se, že do roku 2028 dosáhne velikosti **348,9 milionu USD.**`

Hlavní trendy ve vývoji médií zahrnují zlepšení a rozšíření využívání multimodálních modelů, jako je kombinace LLM a difuzních modelů, schopnost převádět 2D obrázek na 3D pro **realistickou generaci videa z jediného snímku**, rychlejší a přizpůsobitelné metody pro tvorbu upraveného videa v reálném čase a modely, které potřebují méně vstupních dat pro přizpůsobení výsledků, jako je například syntetické audio, které dokáže zachytit charakteristiky člověka z několika sekund referenčních dat.

Tyto trendy **směřují k dalšímu rozvoji v oblasti deepfake útoků**. Hlavní trendy v detekci a autentizaci směřují k vzdělávání, zdokonalení detekce a zvýšenému tlaku ze strany AI komunity na používání technik autentizace médií.

# 7. Jak zkoumat a odhalovat deepfake?

Základní postup samotného **zkoumání a analýzy deepfake** obsahu je následující:

1. Vytvořte kopii média před jakoukoliv analýzou.
2. **Zkontrolujte zdroj** (tj. zda je organizace nebo osoba důvěryhodná) média před vyvozením závěrů.
3. Reverzní vyhledávání obrázků jako je TinEye, Google Image Search a Bing Visual Search může být velmi užitečné, pokud jsou média kompozicí obrázků.
4. Vizuální/audio zkoumání – nejprve se na obsah pečlivě podívejte a poslouchejte, protože mohou být **zřejmé známky manipulace**. Hledejte fyzikální vlastnosti, které by nebyly možné, jako jsou nohy nedotýkající se země, přítomnost zvukových filtrů, jako je šum přidaný pro zamaskování, hledejte nesrovnalosti.
5. Nástroje pro zkoumání metadat mohou v některých situacích poskytnout další klíčové informace. Odstranění některých metadat naznačuje potenciální manipulaci, ale **vyžaduje další vyšetřování**. Absence metadat může naznačovat, že média byla získána prostřednictvím sociálních médií nebo jiného procesu, který automaticky odstraňuje informace.

Pokročilejší doporučení již počítají s následujícími kroky:

1. Kompletní kontrola **specializovaným softwarem** pro ověření obrazových bodů, odrazů a stínů.
2. Zkoumání založené na kompresi a použití nástrojů navržených pro hledání **artefaktů komprese**.
3. Zkoumání založené na obsahu (pokud je to vhodné) a použití nástrojů navržených pro hledání konkrétních manipulací. Například shoda s výstupy z nástrojů dostupných na GitHubu.

Praktické rady pro **zkoumání a odhalování deepfakes** přibližuje Veronika Batelková, CEO projektu **Zvol si info**: „Na problematiku deepfake videí upozorňujeme už 7 let. A jedna věc zůstává stále stejná - kdykoliv se ptáme na nějakém z našich workshopů, zda lidé vědí, co je deepfake, vždy se přihlásí minimálně jeden člověk, který o tomto fenoménu nikdy neslyšel. Ukazujeme pak příklady a lidé žasnou, že je něco takového možné - a žasnou i u starších videích, které jsou ve srovnání s těmi dnešními vlastně k smíchu. Z toho tedy plyne jedna základní rada - zvyšování povědomí o existenci tohoto nebezpečí, protože ačkoliv nám může připadat, že toto téma je už známé, tak pro mnohé opravdu není.

Tipy pro odhalování deepfakes, co Zvol si info doporučovalo historicky:

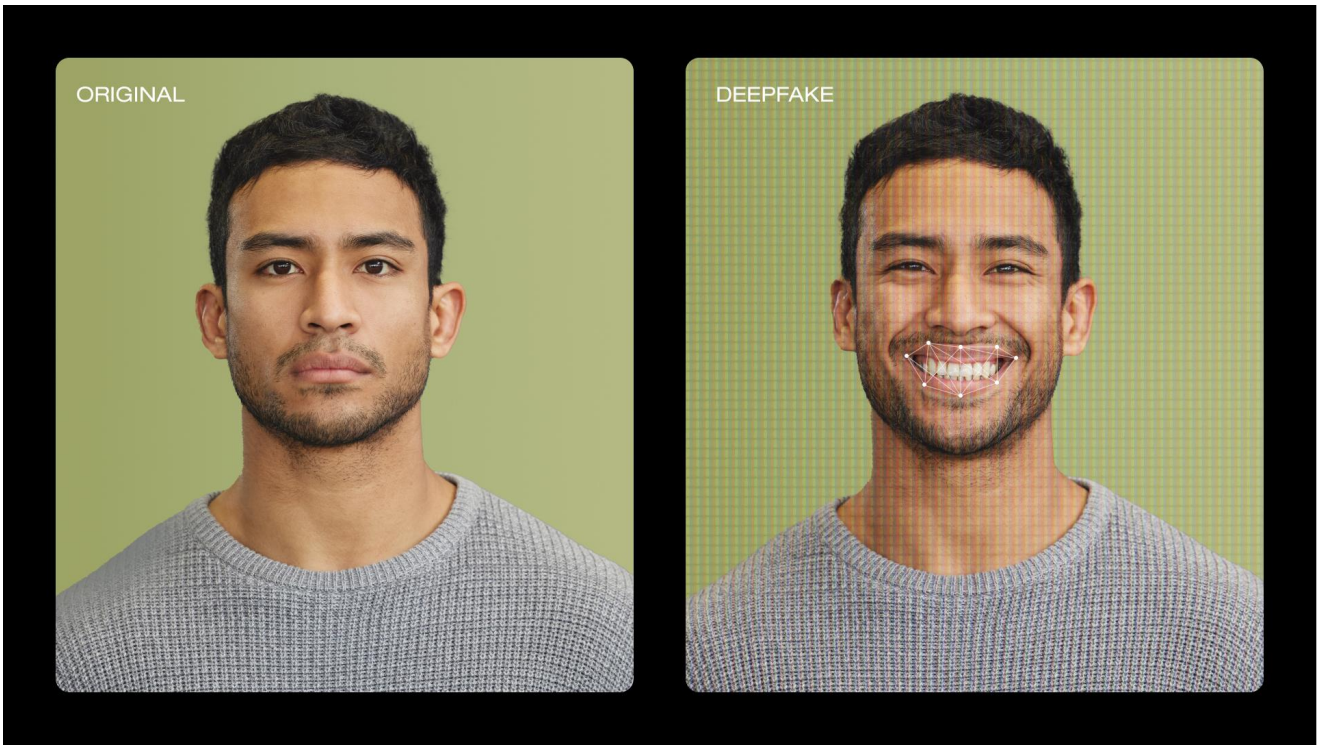
- **Oči** - nepřirozené mrkání nebo jeho úplná absence
- **Ústa** - synchronizace se zvukem, občasný detailní nesoulad mezi pohybem rtů a mluveným slovem
- **Kůže** - nepravidelnosti v textuře
- **Uši** - rozostřené okraje, nepřirozeně umístěny
- **Obecně vzhled** - v některých případech mohou být některé části obličeje nebo hlavy nepřesně sladěny, což způsobuje nesoulad v celkovém vzhledu
- **Chybějící mikrovýrazy** - obličej je bez emocí
- **Pozadí** - postava zvláště „vystupuje“ z pozadí, pozadí je nepřirozené, v nesouladu
- **Nepřirozené pohyby nebo chování** - recyklovaná gesta rukou

Tyto tipy se můžou zdát možná zastaralé, protože se technologie stále zlepšuje, i tak je ale užitečné tyto tipy a triky znát, protože někteří lidé stále mohou používat ne tolik pokročilý software a tak mohou i tyto doporučení pomáhat v boji proti deepfakes.

„Pouze zmíněné tipy však nestačí," dodává Batelková. Je důležité především přemýšlet a ptát se na otázky typu:

- Je chování osoby ve videu **konzistentní** s jejím známým chováním?
- Jaký je kontext dané události?
- A je kontext videa věrohodný?
- **Odkud video pochází?**
- Jak důvěryhodný je zdroj?
- Jaký může být účel a motivace zdroje?
- Existuje potvrzení nebo popření videa (nebo jeho obsahu) od důvěryhodných zdrojů?
- Je prezentace informací **objektivní nebo subjektivní**?

Ačkoliv tedy mohou být výše zmíněné klasické tipy v některých případech užitečné, spíše je důležité **vést zaměstnance ke kritickému myšlení**, než jim dávat jasný návod, jak deepfake rozpoznat.

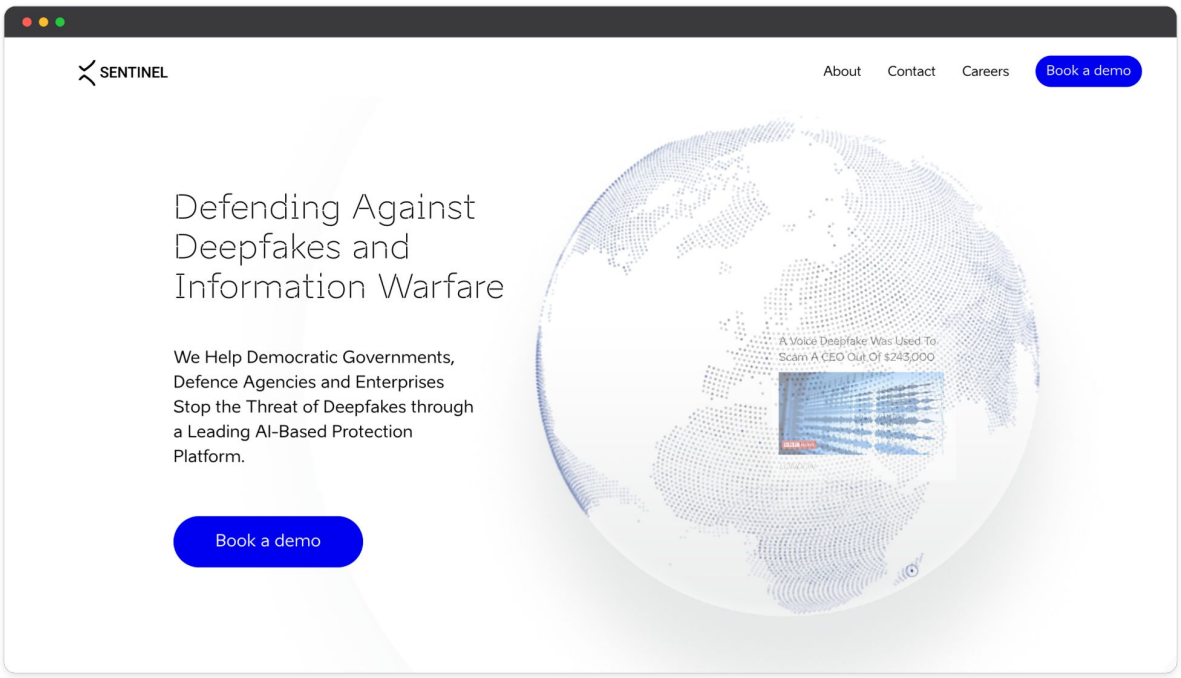


# 8. Software pro detekci deepfake obsahu

Sofistikovaná videa vygenerovaná umělou inteligencí dokážou přesvědčivě napodobit skutečné lidi. Jak však pokročila technologie deepfakes, pokročily i nástroje a techniky určené k jejich detekci. Do tohoto manuálu jsme vybrali 5 nástrojů a technik pro detekci deepfake, které jsou dnes k dispozici.

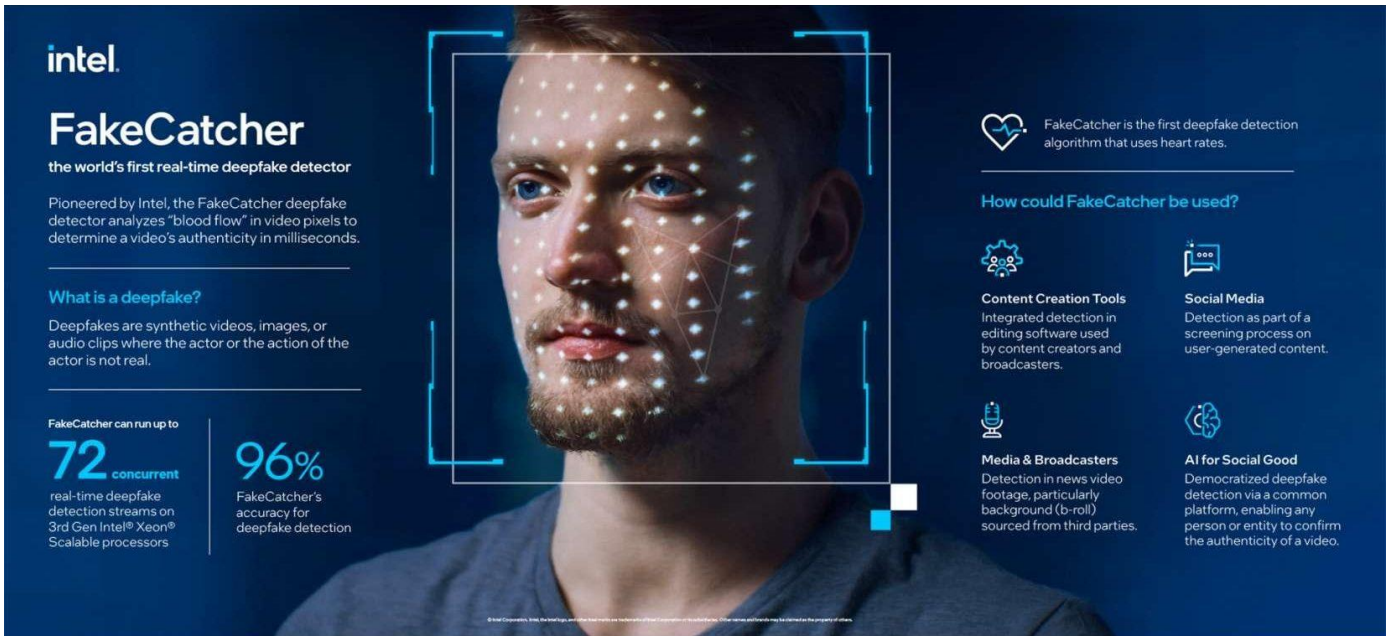
## Sentinel

Sentinel je **přední platforma založená na umělé inteligenci**, která pomáhá vládám, agenturám a podnikům bojovat s deepfakes. Technologii Sentinel je používána předními organizacemi v Evropě. Systém funguje tak, že umožňuje uživatelům nahrávat digitální média prostřednictvím jejich webových stránek nebo API, která jsou pak **automaticky analyzována**. Systém určí, zda je médium deepfake nebo ne, a poskytne vizualizaci manipulace. Systém poskytuje podrobnou zprávu o svých zjištěních včetně vizualizace oblastí médií, které byly změněny.



## FakeCatcher

Intel vyvinul software založený n a sledování prokrvení obličeje lidí na videu, k čemuž využívá umělou inteligenci schopnou učení se. Na základě rozpoznávání jednotlivých pixelů pak dokáže rozpoznat, **zda na videu hovoří skutečný člověk**, či se jedná o digitálně vytvořeného dvojníka. Software se zaměřuje na detekci skutečných věcí a počítá se s možností jeho implementace do aplikací.

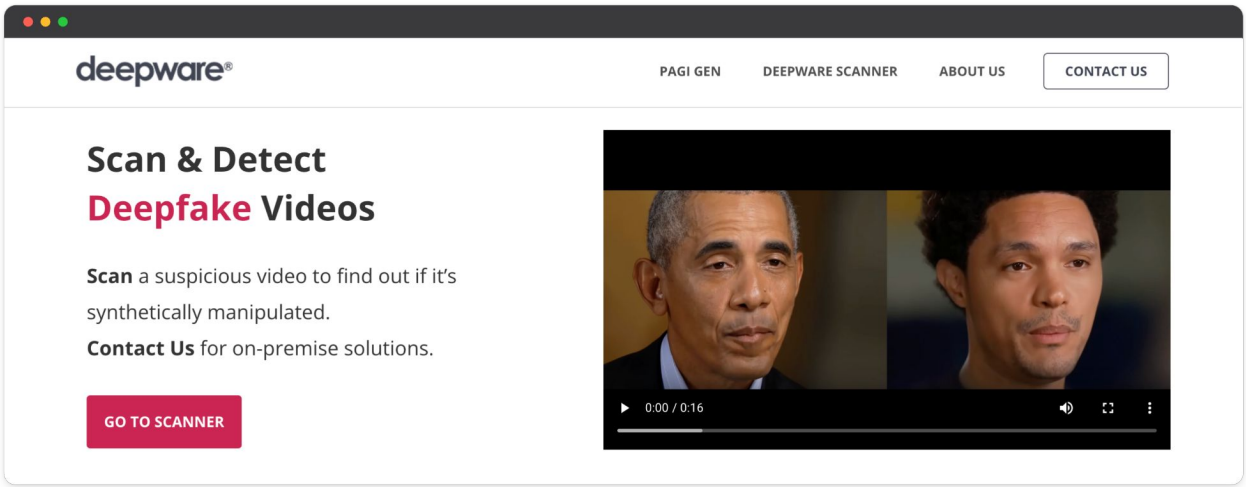



## DuckDuckGoose

DeepDetector je síť vycvičená k rozpoznání tváří generovaných nebo manipulovaných AI. DeepDetector se specializuje na odhalování obsahu generovaného umělou inteligencí, což z něj činí vhodný nástroj v boji proti deepfake. DeepDetector detekuje deepfake tváře **ve videích a obrázcích s přesností 95 %**. Technologie nejen klasifikuje vstup jako falešný, ale také vysvětluje důvody svého rozhodnutí pomocí aktivačních map. Mezi nabízenými řešeními je také AI Voice Detector, který funguje také na češtinu.

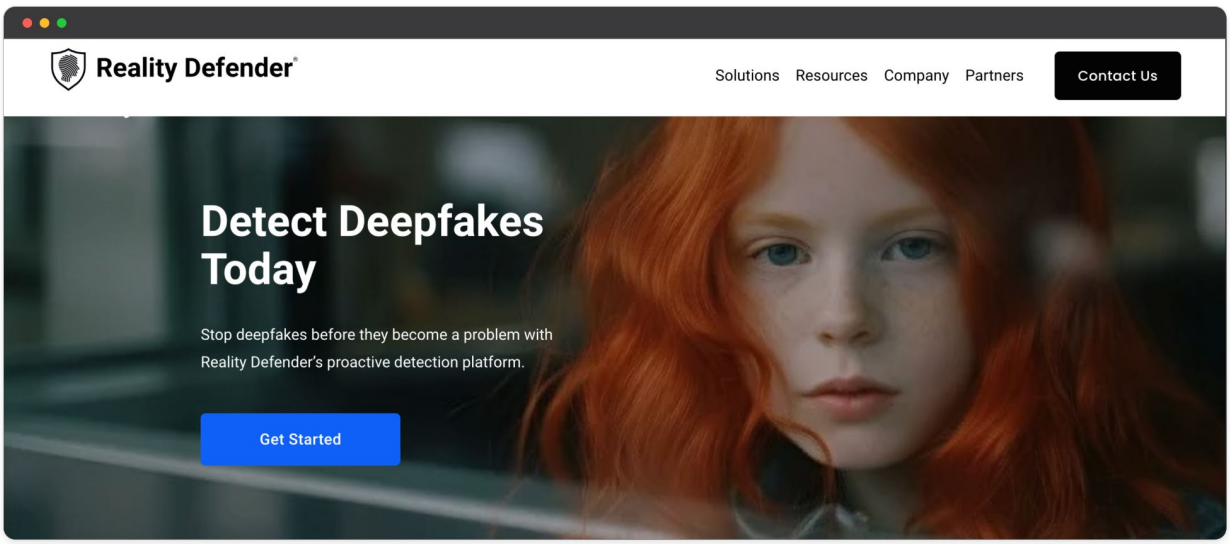
## DeepWare AI

Open-source nástroj s aktivní komunitou, která se zajímá o pokrok v úsilí o detekci deepfake obsahu. DeepWare má přístup ke stále rostoucí knihovně **různorodého video obsahu**, aby bylo zajištěno, že detektor spolehlivě rozpozná syntetická média.




## Reality Defender

Jeden z velmi nadějných startupů vyvíjejících nástroj pro **odhalování deepfakes** a další obsah generovaný AI. Díky komplexnímu a proaktivnímu skenování, praktickým výsledkům a podrobným zprávám pomáhá podnikům a organizacím zůstat v bezpečí.



# 9. Desatero obranné strategie pro české firmy

Organizace mohou podniknout řadu kroků k přípravě na identifikaci, obranu a reakci na hrozby deepfake. Společně s **advokátní kanceláří Legitas** jsme vytvořili **desatero obranné strategie** pro české firmy v boji proti deepfake útokům.

1. **Školení a osvěta zaměstnanců:** Pořádejte pravidelná školení pro zaměstnance o tom, co jsou deepfakes a jak je identifikovat. Učte je na příkladech a ukazujte, jak rozpoznat podvržený obsah.
2. **Zavedení interních bezpečnostních politik:** Stanovte jasná pravidla pro sdílení a ověřování informací uvnitř firmy. Zahrněte pravidla pro komunikaci na sociálních sítích a prostřednictvím e-mailu.
3. **Využití softwaru pro detekci deepfake:** Investujte do technologií a softwaru, které jsou schopny identifikovat deepfake obsah včetně AI nástrojů pro rozpoznávání upravených videí a fotografií.
4. **Zabezpečení digitální identity:** Používejte silné ověřovací metody a dvoufaktorovou autentizaci pro všechny důležité účty a systémy. Zajistěte, aby digitální identity zaměstnanců byly chráněny.
5. **Vytvoření reakčního plánu na incidenty:** Mějte připravený plán, jak reagovat v případě, že se vaše firma stane cílem deepfake útoku. Zahrňte kroky pro komunikaci s médii, zákazníky a ostatními zainteresovanými stranami.
6. **Pravidelné monitorování online prostředí:** Sledujte aktivně sociální média a internet pro možné náznaky deepfake obsahu souvisejícího s vaší firmou nebo zaměstnanci.
7. **Síťová a informační bezpečnost:** Ujistěte se, že vaše firemní sítě a databáze jsou bezpečné a chráněné před neautorizovaným přístupem, který by mohl vést k úniku dat použitelných pro vytváření deepfake materiálů.

8. **Spolupráce s externími experty a bezpečnostními agenturami:** Vytvořte spojení s externími odborníky na kybernetickou bezpečnost pro sdílení informací a nejlepších praktik.
9. **Průběžná revize a aktualizace strategie:** Pravidelně aktualizujte a přizpůsobujte svou obrannou strategii v závislosti na nejnovějších trendech a vývoji v oblasti deepfake technologií.
10. **Právní připravenost a poradenství:** Mějte připravený právní tým specializující se na kybernetickou kriminalitu a ochranu duševního vlastnictví pro případ potřeby právních kroků nebo poradenství. Specializované konzultace s odborníky a kvalitně položené základy vás budou doprovázet léta. Z právního hlediska samozřejmě platí – smlouvy, smlouvy a zase smlouvy. Máte například ošetřenou odpovědnost zaměstnanců v případě, kdy „naletí“ na deepfake, nebo ho dokonce nevědomky budou sdílet a ovlivní tak negativně chod firmy?

## Co dělat, když prevence nezafungovalo?

Je důležité rychle identifikovat a definovat povahu a rozsah problému. To zahrnuje určení, zda došlo k narušení dat, **zneužití deepfake technologie**, nebo k jinému incidentu souvisejícímu s AI. Po nezbytné identifikaci je klíčové neprodleně informovat všechny zúčastněné strany včetně partnerů, zaměstnanců a případně i státních orgánů. **Transparentnost a rychlá komunikace** mohou pomoci minimalizovat škody. Dále se samozřejmě obraťte na právníky (ideálně se specializací na AI a kybernetické právo), aby byla minimalizována jakákoliv další rizika a škody.

Co na to EU AI Act? U generativní AI nařizuje **EU AI Act**, aby byli jednotlivci vždy informováni při interakci s AI a specifický AI obsah (včetně deepfakes) byl označen a musí být detekovatelný.

# 10. Osvědčené postupy z praxe

Pojďme se v samotném závěru tohoto manuálu podívat na „**best practices**“ z firem, které již v boji proti deepfake realizují konkrétní kroky. Pochopitelně ne každý postup bude vhodný pro každou firmu, nicméně je zajímavé podívat se na to, jak deepfake útoky řeší firmy v praxi.

- Měli byste prosazovat firemní kulturu, ve které zaměstnanci mohou zpochybňovat legitimitu informací a **hlásit jakoukoli podezřelou aktivitu**.
- Uveďte zaměstnancům příklady požadavků, které jsou abnormální nebo se vymykají běžným firemním postupům. Zaměstnanci by se nikdy neměli bát o těchto otázkách mluvit. Zaveďte pro hlášení **přesně definovaný** komunikační kanál.
- Mějte směrnici, která nařizuje komunikovat jen prostřednictvím přesně definovaných platforem.
- Velmi detailně prověřujte **zadávání platby novému dodavateli** nebo nový IBAN a důkladně si prověřujte osobu, která platbu požaduje poprvé.
- Zaveďte specifické fráze či gesta pro vysoké představitele firmy, které jsou používány během videohovorů a slouží k ověření jejich identity.
- Provádějte **kontrolované deepfake útoky** v rámci firemního prostředí jako součást školení zaměstnanců. Tím se zvyšuje povědomí a schopnost zaměstnanců rozpoznat podvodné techniky, které používají deepfake.
- Zavedení pravidla, že vždy při důležitých rozhodnutích, zejména těch finančních nebo strategických, musí dojít k osobnímu ověření **v offline světě**.
- Sestavení specializovaného týmu bezpečnostních expertů, kteří se věnují detekci a reakci na deepfake útoky.

# 11. Vzor interní bezpečnostní směrnice

**Úvod**
Tato politika definuje interní postupy a protokoly naší firmy pro ochranu proti rizikům spojeným s deepfake technologiemi. Je zásadní, aby všichni zaměstnanci dodržovali tyto směrnice.

**1. Školení zaměstnanců**
Všichni zaměstnanci se musí zúčastnit pravidelných školení o deepfake technologiích. Zaměstnanci budou informováni o metodách identifikace podezřelého obsahu a procesu nahlášení potenciálních hrozeb.

**2. Standardizace komunikačních kanálů**
Všechna interní a externí komunikace se musí provádět prostřednictvím schválených platform. Jakákoli komunikace obsahující citlivé nebo důvěrné informace mimo tyto kanály je zakázána.

**3. Bezpečnostní protokoly pro autorizaci plateb**
Všechny požadavky na platby vyžadují dvoustupňové ověření přes schválené kanály. Před autorizací platby je nutné ověření identit zúčastněných osob prostřednictvím telefonického hovoru nebo videohovoru.

**4. Zabezpečení digitální identity**
Každý zaměstnanec musí používat silná hesla a aktivovat dvoufaktorovou autentizaci pro všechny firemní systémy.

**5. Reakční plán na incidenty**
V případě zjištění deepfake útoku se aktivuje krizový tým pro řešení situace. Máme připravený komunikační plán pro média a zainteresované strany v případě deepfake incidentu.

### 6. Monitoring

Používáme softwarové řešení pro detekci deepfake obsahu a pravidelně monitorujeme naši online prezentaci. Jakékoli podezřelé aktivity musí být okamžitě nahlášeny bezpečnostnímu týmu.

### 7. Právní připravenost

Náš právní tým je informován o deepfake rizicích a je připraven zasáhnout v případě právních kroků.

### 8. Externí spolupráce a poradenství

Udržujeme kontakty s externími bezpečnostními agenturami a odborníky pro sdílení informací a nejlepších praktik.

### 9. Revize politiky

Tato politika bude pravidelně revidována a aktualizována podle nejnovějších trendů a vývoje.

### 10. Reportování a dokumentace

Zaměstnanci jsou povinni reportovat všechny podezřelé aktivity a uchovávat záznamy pro možná vyšetřování.

---

Tento vzor interní bezpečnostní směrnice by měl být **upraven a přizpůsoben** specifickým potřebám a požadavkům vaší organizace. Za jeho implementaci neneseme odpovědnost.

# Závěr

**Česká asociace umělé inteligence** vám děkuje za projevený zájem a odhodlání vzdělávat se v kriticky důležité oblasti. Vaše rozhodnutí přečíst si tento manuál je jasným důkazem vašeho závazku chránit vaši organizaci před neustále se vyvíjejícími hrozbami spojenými s umělou inteligencí, mezi které AI deepfakes bez diskuze patří.

Je pravdou, že **v roce 2024 budou AI deepfakes představovat významnou výzvu**, ale nezapomeňte, že vaše proaktivní kroky a ochota se učit a adaptovat jsou nejsilnějšími nástroji, které máte k dispozici. Tento manuál vám poskytl nejen cenné informace a strategie, ale také důležitý základ, na kterém můžete stavět vaše další obranné mechanismy.

Věříme, že každý může účinně čelit výzvám, které deepfake technologie představují. Avšak jak už bylo řečeno, nezbytnou součástí vaší strategie musí být **neustálá bdělost**. Kybernetický svět se neustále vyvíjí a s ním i metody útoků. Je tedy klíčové zůstat informováni.

Jménem České asociace umělé inteligence vám přejeme mnoho sil ve vašich snahách o bezpečnější a odolnější firemní prostředí. Společně můžeme čelit hrozbám a zároveň **využívat pozitivní potenciál**, který AI nabízí pro naše podnikání i společnost jako celek.

Budeme velmi rádi za **zpětnou vazbu** a vaše nápady, jak tento manuál ještě vylepšit. Neváhejte se nám ozvat na e-mailovou adresu **info@asociace.ai**.

Nezapomeňte nás také **sledovat na sociální síti LinkedIn**, kde pravidelně přinášíme novinky ze světa umělé inteligence a aktivit naší asociace.

# Výzva ke sdílení

Dovolte nám připojit ještě jednu důležitou poznámku k tomuto manuálu. Rádi bychom zdůraznili, že tento manuál je **volně šiřitelný** a může být sdílen dalším zájemcům, ať už se jedná o jednotlivce, firmy, nebo jiné organizace. Věříme, že sdílením tohoto manuálu můžeme rozšířit povědomí a vzdělání o AI deepfakes mezi širší komunitu.

Nicméně je důležité upozornit, že přestože podporujeme široké sdílení tohoto materiálu, **manuál nesmí být jakkoliv editován nebo modifikován**. Toto omezení je zde z důvodu udržení integrity a přesnosti informací, které obsahuje.

VERZE: **1.0** | Publikováno: 27. prosince 2023