SECURITY DAYS

DIRBTINIS INTELEKTAS: PLONA RIBA TARP PATOGUMO IR RIZIKŲ

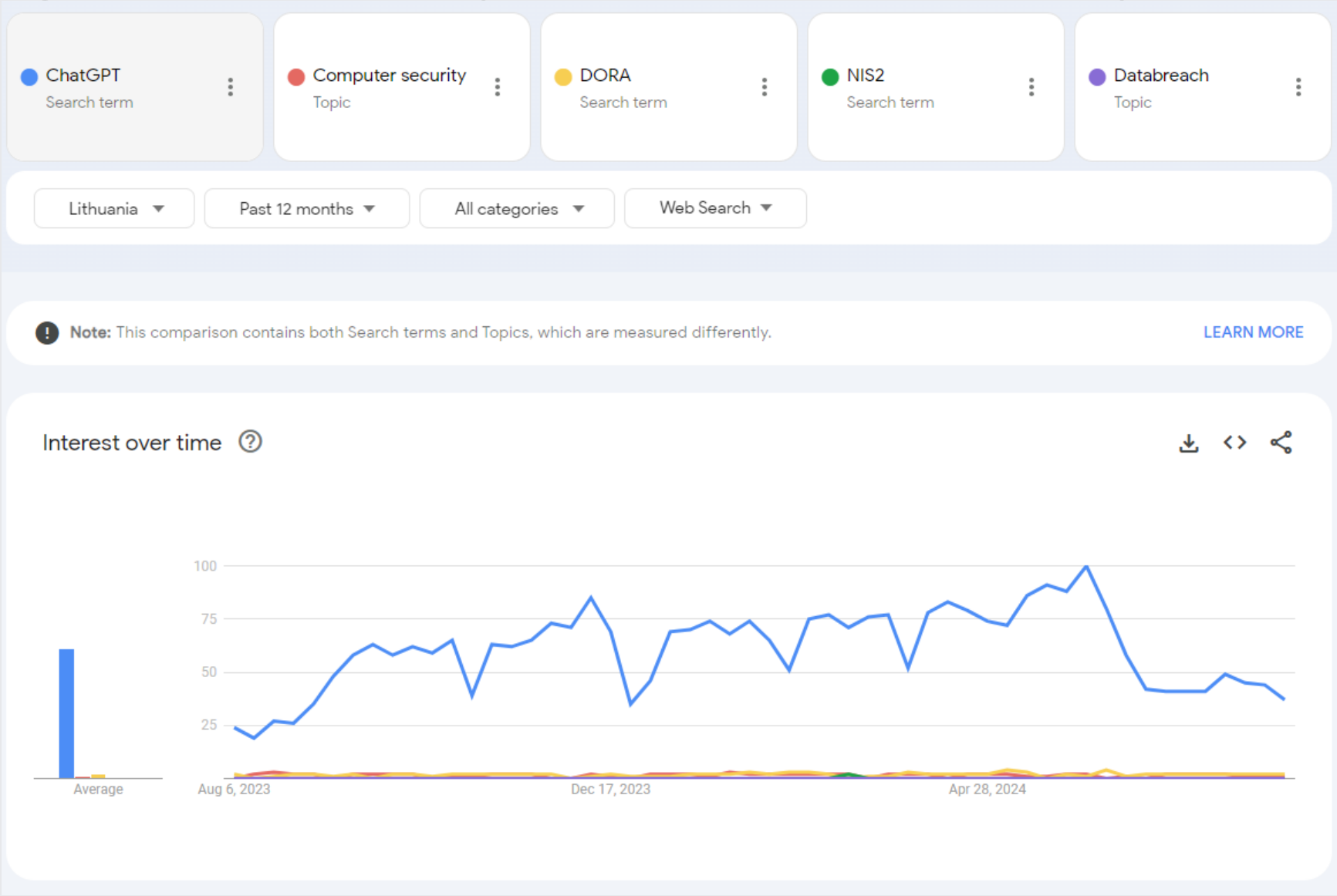
Edita Pulkauninkė - Squalio Lietuva direktorė

Rugsėjo 5 d. 2024

KIBERNETINIO SAUGUMO KONFERENCIJA

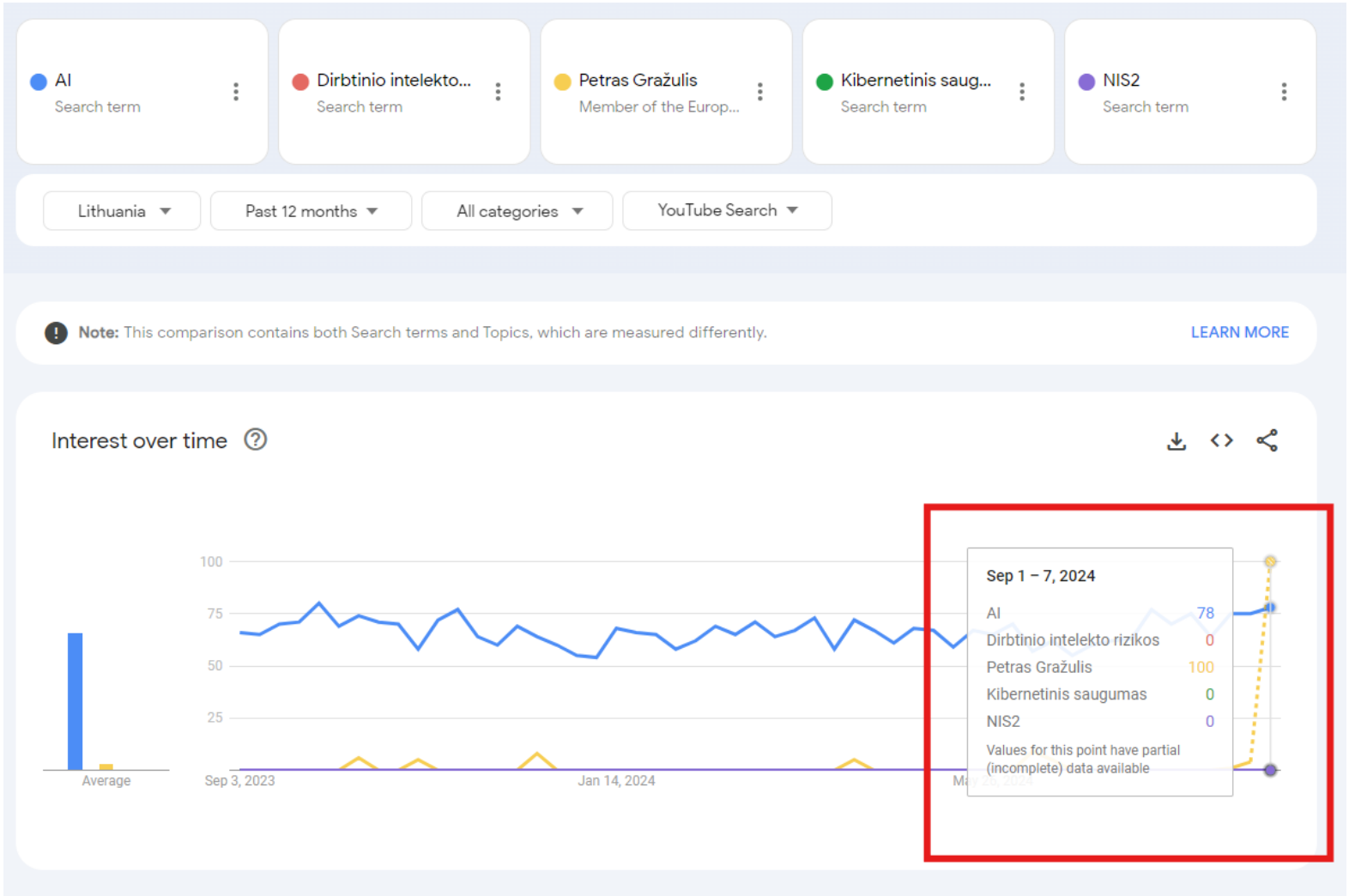
esd.eset.lt

AI ”banga”: pastarojo meto tendencijos

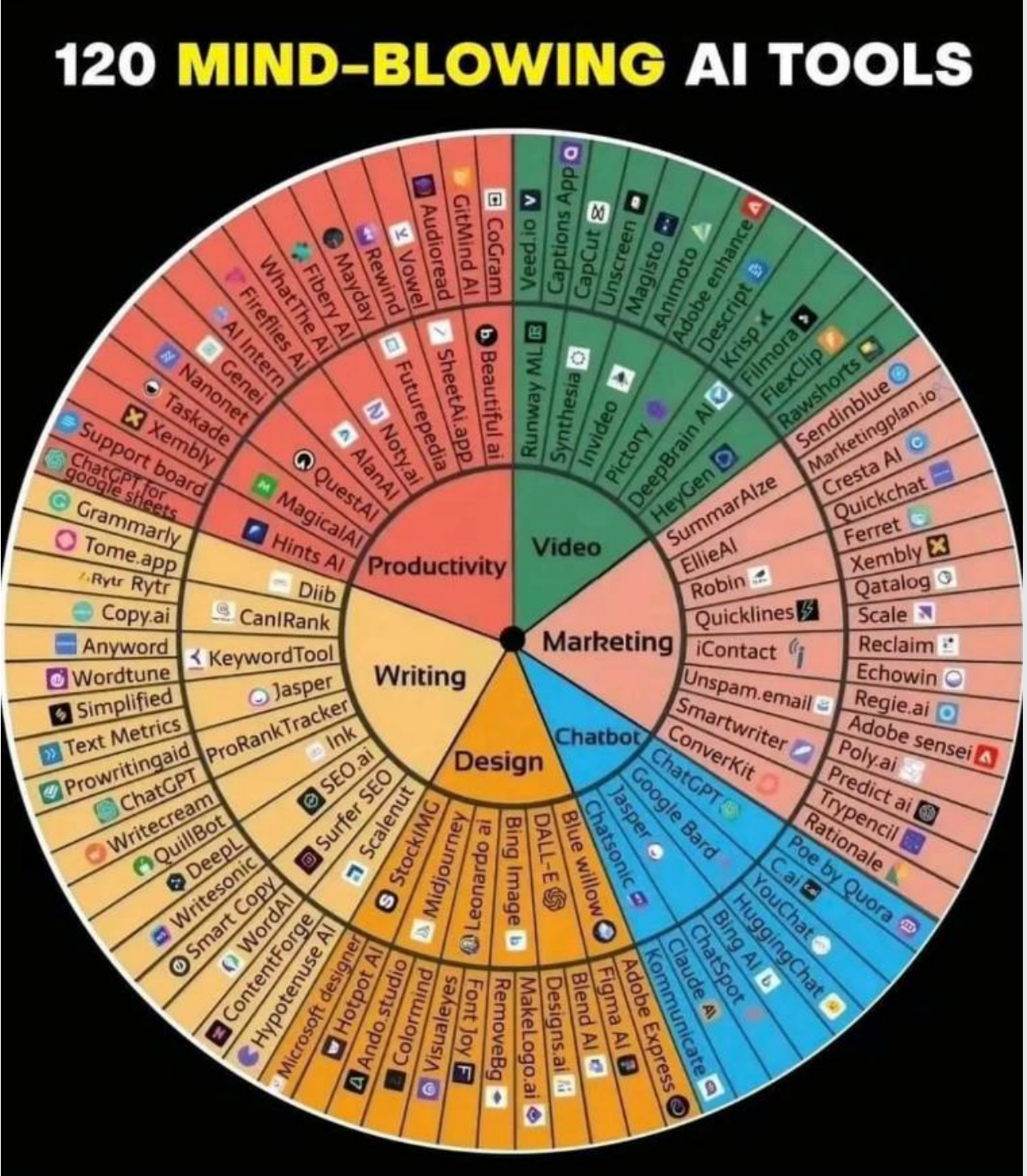


squalio

AI ”banga”: pastarojo meto tendencijos



squalio



GEN AI – ne tik ChatGPT; Google Bard ar Copilot tai kur kas daugiau nei galima tikėtis…

squalio

Kitaip sakant...

squalio

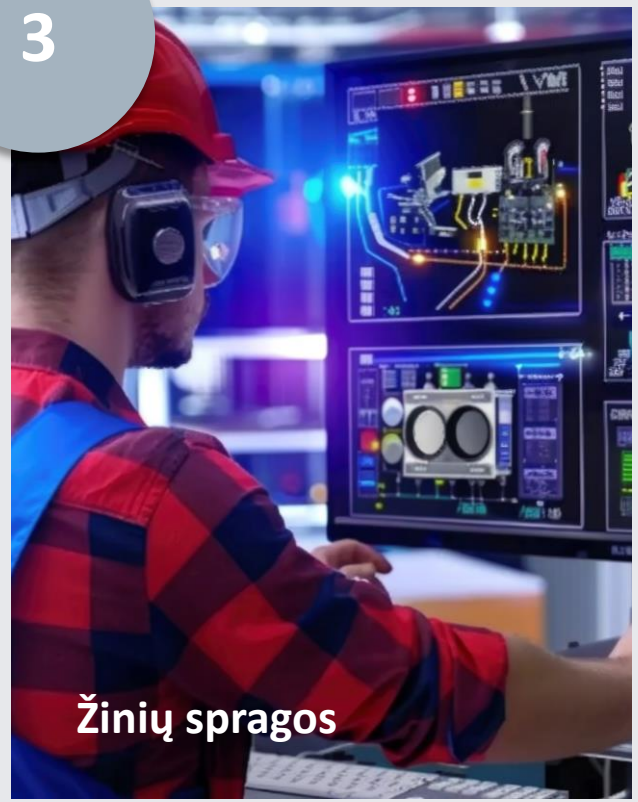
squalio

Nuolat kintančios grėsmės

1

2

3



squalio

DI ”banga”: pastarojo meto tendencijos

squalio

AI rizikų duomenų bazė

- DI rizikų duomenų bazė sugaudo **700+** rizikų faktorių iš **43** egzistuojančių schemų;
- *The Causal Taxonomy of AI Risks* klasifikuoja **kaip, kada,** ir **kodėl** šios rizikos atsiranda;
- *The Domain Taxonomy of AI Risks* klasifikuoja skirtingas rizikas į septynias kategorijas (pvz., “Dezinformacija”) ir 23 subkategorijas (pvz. “Klaidinga arba klaidinanti informacija)

AI Risk Database								High-level Causal Taxonomy			Mid-level Domain Taxonomy	
Title	QuickRef	Ev_ID	Category level	Risk category	Risk subcategory	Description	Additional ev.	Entity	Intent	Timing	Domain	Sub-domain
TASRA: a Taxonomy and Analysis of	Critch2023	01.02.00	Risk Category	Type 2: Bigger than expected		Harm can result from AI that was not expected to have a	the scope of actions available to an AI technology can be greatly expanded when	2 - AI	2 - Unintentional	2 - Post-deployment	7. AI System Safety, Failures, & Limitations	7.3 > Lack of capability or robustness
TASRA: a Taxonomy and Analysis of	Critch2023	01.03.00	Risk Category	Type 3: Worse than expected		AI intended to have a large societal impact can turn out	Oftentimes, the whole point of producing a new AI technology is to produce a large (usually	2 - AI	2 - Unintentional	2 - Post-deployment	7. AI System Safety, Failures, & Limitations	7.3 > Lack of capability or robustness
TASRA: a Taxonomy and Analysis of	Critch2023	01.04.00	Risk Category	Type 4: Willful indifference		As a side effect of a primary goal like profit or influence,	"All of the potential harms in the previous sections are made more likely if the	1 - Human	2 - Unintentional	2 - Post-deployment	6. Socioeconomic and Environmental	6.4 > Competitive dynamics
TASRA: a Taxonomy and Analysis of	Critch2023	01.05.00	Risk Category	Type 5: Criminal weaponization		One or more criminal entities could create AI to	"It's not difficult to envision AI technology causing harm if it falls into the hands of people	1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
TASRA: a Taxonomy and Analysis of	Critch2023	01.06.00	Risk Category	Type 6: State Weaponization		AI deployed by states in war, civil war, or law	"Tools and techniques addressing the previous section (weaponization by	1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
Risk Taxonomy, Mitigation, and Assessment	Cui2024	02.01.00	Risk Category	Harmful Content		"The LLM-generated content sometimes contains biased,		2 - AI	2 - Unintentional	2 - Post-deployment	1. Discrimination & Toxicity	1.2 > Exposure to toxic content
Risk Taxonomy, Mitigation, and Assessment	Cui2024	02.01.02	Risk Sub-Category	Harmful Content	Toxicity	"Toxicity means the generated content contains rude,		2 - AI	2 - Unintentional	2 - Post-deployment	1. Discrimination & Toxicity	1.2 > Exposure to toxic content
Risk Taxonomy, Mitigation, and Assessment	Cui2024	02.01.03	Risk Sub-Category	Harmful Content	Privacy Leakage	"Privacy Leakage means the generated content		2 - AI	2 - Unintentional	2 - Post-deployment	2. Privacy & Security	2.1 > Compromise of privacy by leaking or correctly inferring sensitive
Risk Taxonomy, Mitigation, and Assessment	Cui2024	02.02.00	Risk Category	Untruthful Content		"The LLM-generated content could contain inaccurate		2 - AI	2 - Unintentional	2 - Post-deployment	3. Misinformation	3.1 > False or misleading information

Source: [The AI Risk Repository (mit.edu)]

Šešėlinis AI:

squalio

Kaip suvaldyti šešėlinį AI?

…pradėkite nuo AI politikos

10% organizacijų turi formalią, visapusišką politiką, reglamentuojančią Gen AI naudojimą

squalio

Trys pagrindinės šešėlinio AI rizikos

Dokumentų dalinimasis

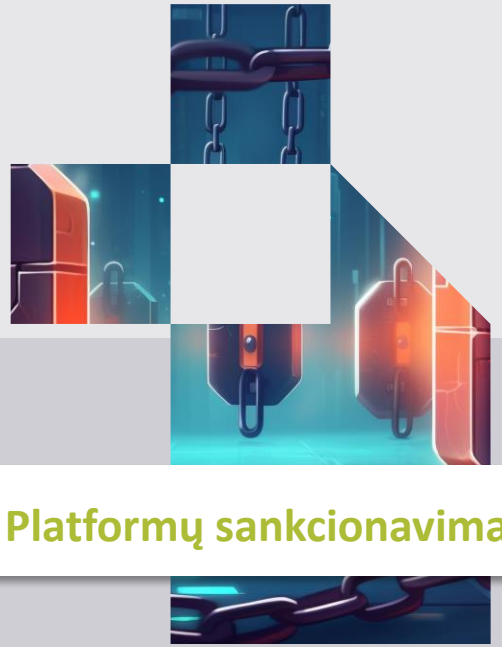
Programinės įrangos diegimas

Įmonės programų diegimas

squalio

Priemonės, galinčios padėti valdyti šešėlinį AI

AI naudojimo politikos



Platformų sankcionavimas

Automatinis prieigos apribojimas

squalio

Prieš diegiant naują AI naudojimopolitiką:

1. **Supraskite, kaip veikia generatyvusis dirbtinis intelektas**
2. **Įvertinkite organizacijos poreikius (*Šešėlinio AI vertinimas*)**
3. **Atlikite rizikos vertinimą**
4. **Patikrinkite esamas IT ir InfoSec politikas**
5. **Supraskite savo techninę aplinką ir jos reikalavimus**

squalio

Šešėlinio AI (kibernetinės saugos) vertinimas

Įvertinti – Išspręsti - Tobulėti

squalio

Šešėlinio DI vertinimas

Tikslinė auditorija

Esate vidutinio dydžio ar didelė įmonė, susirūpinusi dėl kibernetinio saugumo ir privatumo rizikų, kurias kelia nekontroliuojamas generatyviojo DI įrankių, tokių kaip ChatGPT, Bard ir kitų, naudojimas?

Net jei turite įdiegtą programinės įrangos valdymą, dauguma generatyviojo DI įrankių apeina šias sistemas, nes jie veikia naršyklės pagrindu ir nereikalauja vietinės programinės įrangos diegimo.

Jei taip - šis pasiūlymas skirtas Jums, siekiant įvertinti naujausių generatyviojo DI sprendimų naudojimą Jūsų įmonėje.

Privalumai

Matomumas - kas ir kokius generatyviojo AI įrankius naudoja darbuotojai Jūsų įmonėje

Daugiau nei 400 generatyviojo AI paslaugų katalogas (ir šis skaičius auga), įskaitant tik naršyklės pagrindu veikiančias, kurios nėra matomos įprastoms programinės įrangos valdymo sprendimams

Identifikuotų generatyviojo AI paslaugų rizikos klasifikacija pagal bendrus saugumo, atitikties ir teisės kriterijus

Naudojimo duomenys, leidžiantys priimti efektyviausius pirkimo sprendimus bei tobulinti vidines politikos ir instrukcijų procesus

Procesas

- 1 mėnesio procesas
- Microsoft Defender for Cloud Apps diegimas
- Integracija su Microsoft Defender
- Duomenų rinkimas (2 savaitės)

- Projekto rezultatų pristatymas

Rezultatai

Šešėlinio AI vertinimo ataskaita:

- Identifikuotų GenAI įrankių sąrašas
- Sąrašas, kiek vartotojų kuriais įrankiais naudojasi, įskaitant srautą, informacijos mastą bei bendrą interakcijų kiekį
- Kiekvieno įrankio rizikos klasifikacija
- Kiekvieno rizikos lygio naudotojų srauto informacija
- Tiekėjų sąrašas

squalio

Kelios įžvalgos iš vieno vertinimo:

Keletas faktų:

- 65% naudotojų pasitelkia generatyviojo dirbtinio intelekto sprendimus (lyginant su rinkos vidurkiu)
- Identifikuotas 13 GenAI sprendimų naudojimas
- 1Tb bendro srauto, priskiriamo GenAI, nustatymas
- OpenAI ChatGPT GenAI paslaugos – dažniausiai naudojamos kliento įmonėje
- 62% nustatytų GenAI priemonių klasifikuojamos kaip vidutinio arba aukšto rizikos lygio

Pagrindinė rekomendacija:

- Pirmas sprendimas – atlikti bent jau naudojamų programų rizikos vertinimą

squalio

Ataskaitos pvz.:

2.3 Trust score and risk level classification for each ser identified

Table 5. Trust score and risk level classification of identified services

Name	Trust Score	Risk level	Users	Last seen
NaturalReader	2	High	1	YYYY-MM-DD
Lexica	3	High	1	YYYY-MM-DD
ReadSpeaker	5	Medium	6	YYYY-MM-DD
Frase	5	Medium	1	YYYY-MM-DD

2.5 Application usage by user

Table 6. Application usage by user

Application	Application risk level	Users	Last seen
NaturalReader	High	email.address@client.domain	YYYY-MM-DD
Lexica	High	email.address@client.domain	YYYY-MM-DD
Wordtune	Medium	email.address@client.domain	YYYY-MM-DD
ReadSpeaker	Medium	email.address@client.domain	YYYY-MM-DD
ReadSpeaker	Medium	email.address@client.domain	YYYY-MM-DD

2.2 Vendor geographies

Countries listed for vendor headquarters. Locations of service endpoints may

Figure 1. Geolocation of discovered apps (vendor HQ)

Table 4. Vendor HQ locations

Country name	Number of Apps
United States	7
Australia	1
Israel	1
Canada	1
United Kingdom	1
Japan	1

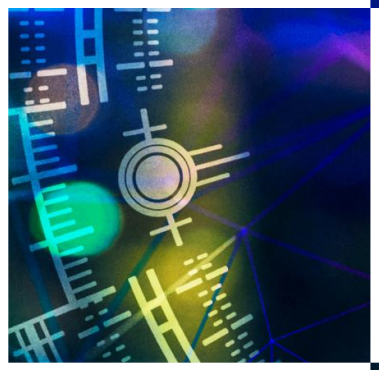
squalio

AI manipuliavimo rizikos

- Renkantis AI modelį, prioritetą teikite saugumui
- Rinkitės patikimus tiekėjus
- Jei Jūsų techninės kompetencijos ribotos, venkite nemokamų įrankių
- Taikykite nuolatinę įrankio ir sistemos stebėseną
- Prašykite programuotojų ištestuoti pasirinktą

modelį

- Būkite pasiruošę saugumo incidentams



squalio

Mes galime padėti

Licencijavimo klausimu

Access a vast array of software solutions from leading vendors, all in one place. Squalio covers productivity tools, CRM software, project management, and ERP systems.

Microsoft Solutions Partner

Google Cloud Partner

Adobe Gold Reseller

Kibernetinis saugumas

Our solutions go beyond tools and focus on continuously raising awareness among your employees.

- Cybersecurity Testing & Audits
- NIS2 Compliance Managed Service
- Zero Trust Maturity Assessment
- SIEM (Security Information and Event Management) Solutions
- User Cybersecurity Training & Social Engineering
- Information Security Policy Creation
- GDPR (General Data Protection Regulation) and other Industry Standards Compliance Assessments
- Red Team-based Cybersecurity Assessments
- Cybersecurity Specialist as a Service

Infrastruktūra

Whatever your infrastructure needs, be they at the data center or at the edge, Squalio has the experience you need to build and support IT infrastructure -even in the most demanding and complex projects.

- IT Infrastructure Assessments and Audits
- IT Infrastructure Development
- IT System Migration
- IT System Monitoring
- IT System Maintenance
- Cloud Solutions
- Productivity Platform Migration & Maintenance

squalio

Mes galime padėti

AI ir Copilot paslaugos

Drive operational innovation with AI solutions, custom-tailored or ready-to-use products to keep your organization at the forefront of technological advancements.

- Copilot HeadStart
- GenAI in a box
- GenAI policy design
- Shadow AI Assessment
- Copilot Security Readiness Assessment
- Azure OpenAI Deployment
- Azure OpenAI Healthcheck
- Prompt engineering training
- GenAI managed policy
- Copilot Extended
- Azure OpenAI managed service
- GenAI safety managed service
- Custom RAG and knowledge graph solutions

ITAM

ITAM is not just about software, it is about hardware, cloud, SaaS, and asset management – a comprehensive approach to managing all IT assets, from hardware to software to cloud. We use best-in-class tools to make managing your business IT more efficient and effective. We help you take control of IT spending, improve compliance, and reduce risk.

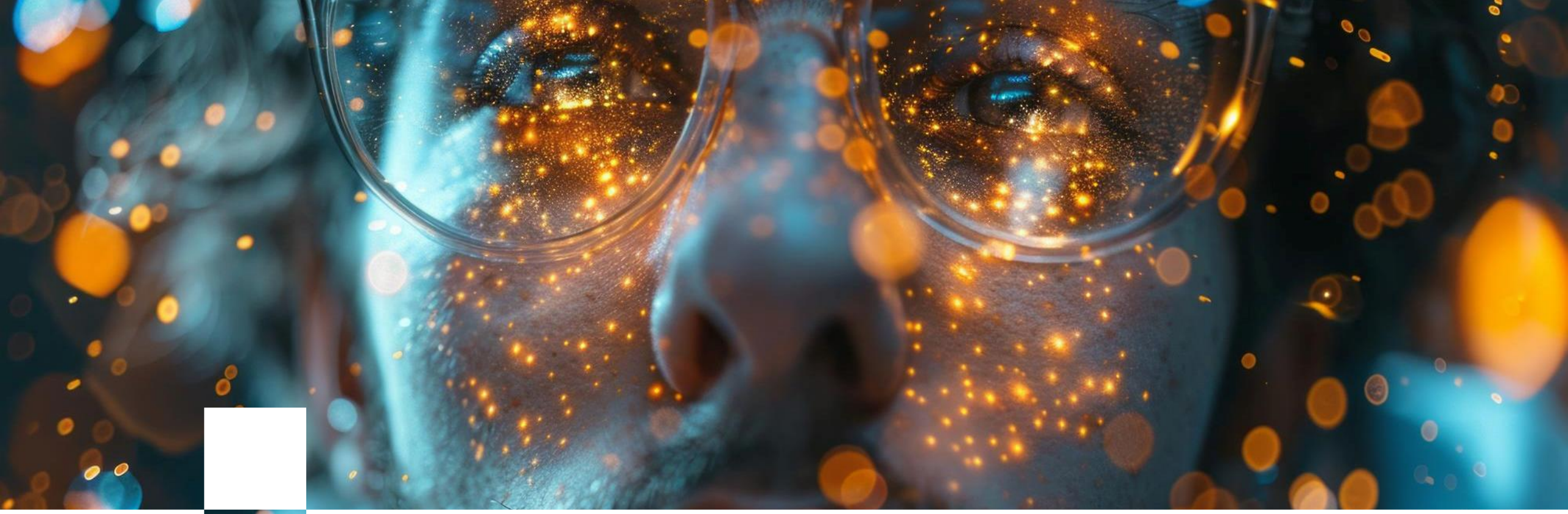
- Consulting Services
- Deep Vendor Expertise
- Technical Services & Tools
- Managed Services

GDPR klausimai

Leverage the full potential of your data with Squalio's specialized services utilizing Microsoft Fabric, as well as be sure your business is fully GDPR compliant.

- GDPR trainings
- GDPR compliance assessments
- Data Protection Officer as a service
- Microsoft Fabric

squalio



“**Išlieka ne stipriausios ir net ne protingiausios rūšys**, o tos, kurios geriausiai prisitaiko prie pokyčių.“

Charles Darwin

squalio

Tikiuosi buvo naudinga ;)

Edita Pulkauninkė
Squalio Lietuva vadovė
edita.pulkauninke@squalio.com
+370 611 60711