ITRC
پژوهشگاه ارتباطات و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)

مرجع دانش حملات سایبری مبتنی بر هوش مصنوعی

تاریخ انتشار: شهریور ۱۴۰۴

ITRC

پژوهشگاه ارتباطات
و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)

عنوان گزارش: مرجع دانش حملات سایبری مبتنی بر هوش مصنوعی

کلمات کلیدی: هوش مصنوعی، امنیت سایبری، هوش مصنوعی تهاجمی و هوش مصنوعی تخاصمی

تهیه کنندگان: محمد جاویدنیا، مرجان گودرزی، فائزه غلامرضائی، فاطمه اصغری‌همپا و مائده برنا

راهبر علمی (مجری پژوهش): امیرمنصور یادگاری

گروه پژوهشی: ارزیابی امنیت شبکه و سامانه‌ها (پروژه پژوهشی تدوین طرح مقابله با تهدیدات سایبری بهره‌مند از هوش مصنوعی)

تاریخ نشر: شهریور ۱۴۰۴

ITRC

پژوهشگاه ارتباطات
و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)

خلاصه مدیریتی

سند حاضر، یک مرجع دانش[1] برای آشنایی با مشخصات و تأثیرات حملات سایبری مبتنی بر هوش مصنوعی[2] است. منظور از حملات سایبری مبتنی بر هوش مصنوعی، آن دسته از حملات است که اولاً در فضای سایبری[3] یا با استفاده از فضای سایبری صورت می‌پذیرند و ثانیاً در تمام یا بخشی از مراحل حمله، از قابلیت‌های هوش مصنوعی (اعم از هوش مصنوعی پیش‌بین[4] یا هوش مصنوعی مولد[5]) استفاده می‌کنند. یک حمله مبتنی بر هوش مصنوعی می‌تواند تقویت‌شده با هوش مصنوعی[6] باشد یا اساساً ایجادشده با هوش مصنوعی[7] باشد. هدف[8] یک حمله مبتنی بر هوش مصنوعی، لزوماً یک موجودیت یا سامانه مبتنی بر هوش مصنوعی نیست، اما اگر هدف حمله نیز مبتنی بر (برخوردار از) هوش مصنوعی باشد، علاوه بر حملات تهاجمی[9] مبتنی بر هوش مصنوعی، حملات تخاصمی (متخاصم‌ساز)[10] هوش مصنوعی نیز مطرح می‌شوند که مرجع دانش حاضر، این نوع از حملات را نیز شامل می‌گردد. متخاصم‌سازی سامانه‌ها یا راه‌کارهای هوش مصنوعی با توجه به فراگیر شدن تدریجی استفاده از چنین سامانه‌ها یا راه‌کارهایی در حوزه‌های کاری یا خدمت‌رسانی مختلف، به سرعت به یکی از سرفصل‌های جدّی و نیازمند توجه آتی تبدیل خواهد شد، از این رو، در مرجع دانش حاضر مورد تأکید قرار گرفته است تا آگاهی‌های امنیتی در خصوص بکارگیری آنها در بخش‌های مختلف افزایش یابد.

حملاتی که بهره‌مند از هوش مصنوعی می‌باشند، نسبت به حملات مشابه و رایجی که از هوش مصنوعی استفاده نمی‌کنند، ممکن است قابلیت‌های فراوانی را برای واردسازی صدمات سنگین‌تر و جبران‌ناپذیرتر در نتیجه ارتقاء پنهان‌مانی، پردازش‌گری و خودکارسازی داشته باشند. برای مقابله با

1. Knowledge Handbook
2. AI-based Cyber Attacks
3. Cyber Space
4. Predictive AI
5. Generative AI
6. AI-powered (AI/ML-enabled/enhanced)
7. AI-driven
8. Target
9. Offensive Attacks
10. Adversarial Attacks

این نوع از حملات (از شناسایی تا دفع آنها)، بکارگیری راه‌کارهای معمول در مقابله با حملات مشابه، مؤثر نخواهد بود و در نتیجه به راه‌کارهای مقابله‌ای و امن‌سازی جدید مبتنی بر فناوری‌هایی احتمالاً متفاوت یا پیشرفته‌تر نیاز خواهند داشت که ممکن است از هوش مصنوعی نیز بهره‌برداری کنند. یک نکته این است که متخاصم‌سازی سامانه‌ها یا راه‌کارهای امنیتی نیز یک خطر جدی در آینده نزدیک است که در آن صورت، عامل امنیتی خود به یک عامل ضدامنیتی تبدیل خواهد شد. البته سند حاضر، صرفاً به فهرست کردن سرفصل راه‌کارهای امن‌سازی مفید برای مقابله با حملات پرداخته تا مدیران و کارشناسان بخش امنیت نسبت به کلیتی از اقداماتی که باید برای امن‌سازی در برابر حملات سایبری مبتنی بر هوش مصنوعی صورت پذیرد، مطلع گردند و تمرکز این سند بیشتر بر روی ایجاد یک مرجع دانش برای آشنایی با مشخصات و تأثیرات انواع حملات سایبری تهاجمی و تخاصمی مبتنی بر هوش مصنوعی است.

در سند حاضر، ابتدا شناسنامه انواع حملات سایبری مبتنی بر هوش مصنوعی در دو فصل جداگانه حملات تهاجمی و تخاصمی، آمده است. شناسنامه هر حمله، شامل توضیح مختصری درباره حمله سپس ویژگی‌های کلیدی، تکنیک‌های هوش مصنوعی مورد استفاده، دارایی‌های هدف، آسیب‌پذیری‌ها، سناریو و مراحل اجرا، نشانه‌های احتمالی نفوذ، تأثیرات و تبعات وقوع حمله و راه‌های مقابله و کاهش مخاطره است. پس از دو فصل حاوی شناسنامه حملات، فصل انتهایی مرجع دانش است که شامل جداولی به تفکیک مجموعه‌ای از حوزه‌های مختلف آسیب‌پذیر متصل بر شبکه است که حملات سایبری مبتنی بر هوش مصنوعی می‌تواند دارایی‌های آنها را به خطر بیندازد و مورد آسیب واقع کند و تأثیرات نامطلوبی بر عملکرد کل حوزه ایجاد نماید. حوزه‌هایی که در سند حاضر به آنها پرداخته شده، عبارتند از: حوزه مالی و بانکی، حوزه سلامت، خبرگزاری‌ها و پایگاه‌های خبری، پیام‌رسان‌های همراه، گفتگوگرهای مبتنی بر مدل‌های زبانی بزرگ؛ آنچه در گزارش حاضر آمده، برای آگاهی متولیان حوزه‌های مورد اشاره از حملات سایبری مبتنی بر هوش مصنوعی است که دارایی‌های ملموس و ناملموس تحت مدیریت آنها به خطر می‌اندازد و خطراتی که از طریق آن حوزه‌ها برای دیگر حوزه‌ها یا کاربران آنها ممکن است وجود داشته باشد، موضوع سند حاضر نیست. امید است که انتشار این سند، با افزایش آگاهی‌های امنیتی متولیان حوزه‌های خدمت‌رسان گوناگون به ویژه با فراگیرشدن بکارگیری قابلیت‌های هوش مصنوعی از سوی تهدیدگران امنیتی، برای پایدارسازی و تداوم ارائه خدمات مبتنی بر فضای سایبری کشور مفید و اثربخش باشد.

ITRC

پژوهشگاه ارتباطات
و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)

۱ مقدمه

در طول چند دهه اخیر، انواع مختلفی از سازمان‌ها، از جمله آژانس‌های دولتی، بیمارستان‌ها و مؤسسات مالی، هدف حملات سایبری قرار گرفته‌اند. با گذشت زمان، روش‌ها و پیچیدگی حملات سایبری تکامل یافته‌اند و مهاجمان به طور فزاینده‌ای در بهره‌برداری از آسیب‌پذیری‌های سامانه‌های رایانه‌ای ماهرتر شده‌اند. ظهور هوش مصنوعی (AI)[1] در سال‌های اخیر بُعد جدیدی به حملات سایبری افزوده است و به مهاجمان این امکان را داده تا حملات خود را به صورت خودکار و شخصی‌سازی شده انجام دهند، به گونه‌ای که پیش از این ممکن نبود. پدیده‌ای به نام هوش مصنوعی تهاجمی یا هوش مصنوعی مخرب[2] را معرفی کرده و آن را بهره‌برداری از هوش مصنوعی به منظور انجام یک فعالیت مخرب تعریف می‌کنند.

البته ذکر این نکته ضروری است که هوش مصنوعی در امنیت سایبری یک شمشیر دولبه است. سازمان‌ها می‌توانند از آن برای تقویت دفاع‌های سایبری خود استفاده کنند، اما مجرمان سایبری نیز می‌توانند از هوش مصنوعی برای راه‌اندازی حملات هدفمند با سرعت و مقیاس بی‌سابقه استفاده کنند و از روش‌های تشخیص سنتی فرار کنند. افزایش شیوع حملات سایبری مبتنی بر هوش مصنوعی، ماهیت دوگانه هوش مصنوعی را برجسته می‌کند که می‌تواند هم برای بهبود و هم برای تضعیف امنیت سایبری مورد استفاده قرار گیرد.

با توجه به ادبیات موضوع، بهره‌گیری از هوش مصنوعی در حملات سایبری را در قالب ۳ دسته می‌توان تعریف کرد. هوش مصنوعی تدافعی[3]، هوش مصنوعی تهاجمی و هوش مصنوعی تخاصمی و باید میان آن‌ها تمایز قائل شد. هوش مصنوعی تهاجمی، یا حملات مبتنی بر هوش مصنوعی، به استفاده از هوش مصنوعی به منظور انجام فعالیت‌های مخرب، مانند توسعه حملات سایبری جدید یا خودکار کردن بهره‌برداری از آسیب‌پذیری‌های موجود اشاره دارد.

سامانه‌های هوش مصنوعی تهاجمی می‌توانند به مهاجمان کمک کنند تا حملات خود را به صورت خودکار و در مقیاس بزرگ انجام دهند و به آن‌ها این امکان را می‌دهند که آسیب‌پذیری‌ها را سریع‌تر و با دقت بیشتری مورد بهره‌برداری قرار دهند. هوش مصنوعی تخاصمی یا سوءاستفاده و استفاده

1. Artificial Intelligence (AI)
2. Malicious AI
3. Defensive AI

نادرست از سامانه‌های هوش مصنوعی به نوعی زیرمجموعه‌ای از هوش مصنوعی تهاجمی در نظر گرفته می‌شود و به حملاتی اطلاق می‌شود که از آسیب‌پذیری‌های سامانه‌های هوش مصنوعی بهره‌برداری می‌کند تا باعث شود سامانه هوش مصنوعی به درستی کار نکند. این کار می‌تواند از طریق دستکاری داده‌های ورودی به سامانه هوش مصنوعی یا مسموم کردن داده‌های آموزشی که سامانه با آن‌ها آموزش دیده، انجام شود. هوش مصنوعی تدافعی از یادگیری ماشین (ML)[1] و دیگر تکنیک‌های هوش مصنوعی برای بهبود امنیت و تاب‌آوری سامانه‌های رایانه‌ای و شبکه‌ها در برابر حملات سایبری استفاده می‌کند. جدول ۱ تفاوت‌های کلیدی ۳ دسته را بر اساس منابع مختلف خلاصه می‌کند.

جدول ۱: تفاوت‌ها بین انواع دسته‌های بهره‌گیری از هوش مصنوعی در امنیت سایبری

نوع	هدف	مثال
تهاجمی	به کارگیری تکنیک‌های هوش مصنوعی برای حمله به سیستم‌ها و شبکه‌های کامپیوتری	سرقت و جعل هویت با استفاده از جعل عمیق
تخاصمی	سوءاستفاده و/یا حمله به سیستم‌ها و داده‌های هوش مصنوعی	مسموم کردن داده‌های آموزشی و دستکاری داده‌های ورودی
تدافعی	استفاده از تکنیک‌های هوش مصنوعی برای محافظت از سیستم‌ها و شبکه‌های کامپیوتری در برابر حملات	ضد بدافزار و سیستم‌های تشخیص نفوذ مبتنی بر هوش مصنوعی

باتوجه به ماهیت سند حاضر در شناسایی حملات مبتنی بر هوش مصنوعی، دو دسته هوش مصنوعی تهاجمی و هوش مصنوعی تخاصمی مورد بررسی قرار خواهد گرفت و انواع حملاتی که در مطالعات گوناگون در این دو دسته شناسایی شده‌اند تشریح می‌گردند. سند حاضر در ۴ فصل تهیه و تدوین شده است که در ادامه کلیات مطالب هر فصل بیان می‌گردد:

- فصل اول سند مقدمه‌ای است بر انواع کاربردها هوش مصنوعی در امنیت سایبری و تبیین روند کلی سند و فصل‌های آن.
- فصل دوم سند به شناسنامه حملات مرتبط با هوش مصنوعی تهاجمی می‌پردازد.
- در فصل سوم به شناسنامه حملات هوش مصنوعی تخاصمی پرداخته می‌شود.
- در فصل چهارم نیز به مطالعه موردی حملات در حوزه‌های کاربردی منتخب پرداخته خواهد شد.

1. Machine Learning (ML)

۲ شناسنامه حملات هوش مصنوعی تهاجمی

۲-۱ مقدمه

هوش مصنوعی تهاجمی به استفاده از فناوری‌های هوش مصنوعی، الگوریتم‌ها و سامانه‌ها به شیوه‌هایی اشاره دارد که مضر، خرابکارانه یا غیراخلاقی هستند. این موضوع می‌تواند شامل استفاده از هوش مصنوعی برای اهدافی مانند حملات سایبری، کمپین‌های اطلاعات نادرست، نظارت یا فعالیت‌های دیگری باشد که حریم خصوصی، امنیت یا هنجارهای اخلاقی را نقض می‌کنند.

حملات سایبری مبتنی بر هوش مصنوعی به عنوان یک تهدید اصلی در حال ظهور هستند، زیرا این حملات پیچیده‌تر و متنوع‌تر می‌شوند. هوش مصنوعی می‌تواند مجموعه ابزارهای قدرتمندی را برای مهاجمان سایبری فراهم کند تا تمام انواع حملات سایبری سنتی، از جمله فیشینگ، بدافزارها، حملات به رمزهای عبور را تقویت کند.

به عنوان مثال، تکنیک‌های هوش مصنوعی می‌توانند برای ایجاد بدافزارهایی استفاده شوند که بتوانند به محیط خود تطبیق پیدا کنند، از اعمال خود یاد بگیرند و روش‌های خود را بهبود دهند تا از مکانیزم‌های تشخیص سنتی فرار کنند. بدافزارهای پلی‌مورفیک و متامورفیک نمونه‌های واضحی از بدافزارهای مبتنی بر هوش مصنوعی هستند. این سویه‌ها می‌توانند کد خود را تغییر دهند تا از تشخیص مبتنی بر امضای دیجیتال فرار کنند و در پاسخ به اقدامات مقابله‌ای تکامل یابند که این موضوع چالشی جدی برای دفاع‌های امنیت سایبری به وجود می‌آورد.

علاوه بر این، هوش مصنوعی می‌تواند حملات فیشینگ را مؤثرتر کند، تولید وب‌سایت‌ها و ایمیل‌های جعلی بسیار متقاعدکننده را خودکار کند و احتمال اینکه افراد به آن‌ها فریب بخورند را افزایش دهد. هوش مصنوعی همچنین می‌تواند برای ایجاد حملات فیشینگ هدفمند استفاده شود که افراد یا نهادهای خاصی را هدف قرار می‌دهند. با تحلیل مجموعه داده‌های بزرگ، مهاجمان می‌توانند پیام‌های خود را متناسب با زمینه شخصی یا حرفه‌ای هدف تنظیم کنند که این موضوع احتمال موفقیت آن‌ها را افزایش می‌دهد.

این پیشرفت‌ها پویایی‌های در حال تغییر فضای تهدیدات سایبری را نشان می‌دهند، جایی که هوش مصنوعی به یک ضریب تقویت‌کننده برای هر دو طرف مهاجمان و مدافعان تبدیل شده است. این تهدیدات سایبری مبتنی بر هوش مصنوعی همچنین خطر جدی برای حریم خصوصی و امنیت ایجاد می‌کنند و مکانیزم‌های امنیت سایبری سنتی ممکن است نتوانند با آن‌ها همگام شوند.

در این فصل با بررسی مقالات و تحقیقات بین‌المللی و بر مبنای گزارشات مؤسسات امنیت سایبری و زنجیره کشتار سایبری[1] و اطلس Mitre، ۵۲ حمله از نوع هوش مصنوعی تهاجمی شناسایی

1. Cyber Kill Chain

و برای آن‌ها شناسنامه‌ای تفصیلی مبتنی بر ۸ مؤلفه تدوین گردید. برای هر یک از حملات ابتدا توضیح مختصری بیان شده و سپس موارد زیر مشخص و شرح داده شده است:

- **ویژگی‌های کلیدی:**
  - این بخش به معرفی ویژگی‌هایی هر یک از تهدیدات و حملات می‌پردازد و وجوه تمایز آن را با سایر حملات و تهدیدات اشاره دارد.
- **تکنیک‌های هوش مصنوعی مورد استفاده:**
  - در این بخش برخی از اصلی‌ترین مدل‌ها، الگوریتم‌ها و تکنیک‌های هوش مصنوعی که در بروز حمله ایفای نقش دارند به صورت بالفعل یا باقوه معرفی می‌شوند. بدیهی است با توجه به رویکرد گزارش حاضر که سعی در شناخت حملات هوش مصنوعی تهاجمی دارد این بخش نه تنها به صورت جزئی و فنی که به صورت سطح بالا تدوین گردیده است.
- **دارایی‌های هدف:**
  - این بخش اهداف حملات در یک سازمان، سامانه یا جامعه را معرفی می‌کند.
- **آسیب‌پذیری‌ها:**
  - بیانگر نقاط ضعف داخلی سامانه‌ها، کاربران یا محیط عملیاتی و استقرار آن‌هاست که مهاجمان با سوءاستفاده از آن‌ها اقدام به اجرای حملات خود می‌کنند.
- **سناریو حمله:**
  - در این بخش سناریو و مراحل اجرای اصلی حملات بیان می‌گردد. سناریو و مراحل اجرای بیان شده، پایه‌ای است که با استفاده از آن می‌توان سنایو و مراحل اجرای حمله در حوزه‌های گوناگون کاربردی که هر یک از حملات در آن‌ها معنی دارد را بیان کرد.
- **نشانه‌های احتمال نفوذ:**
  - در این بخش فهرستی از مواردی که در صورت بروز احتمال به وقوع پیوستن و یا در حال وقوع بودن حمله را مشخص می‌کند بیان می‌کند. بدیهی است که ممکن است نشانه‌های مشترکی میان حملات گوناگون وجود داشته باشد.
- **تأثیرات:**
  - این بخش به تبعات، پیامدها و تأثیراتی که در صورت بروز موفق حملات اتفاق می‌افتد اختصاص دارد. این پیامدها می‌تواند در ابعاد سامانه یا محصول، سازمان یا در بُعد فردی و اجتماعی باشد.
- **راه‌های مقابله:**
  - در این بخش مجموعه ای از راهبردها و اقدامات جهت مقابله و کاهش محاطرات هر یک از

حملات بیان می‌گرد. این بخش نیز ممکن است در برخی از مصادیق میان حملات مختلف مشترک باشد.

در ادامه نگاشتی میان حملات ۵۲ گانه شناسایی شده و تاکتیک‌های ۱۴ گانه MITRE انجام شده است که نتایج مربوط به آن در جدول ۲ نمایش داده شده است. ذکر این نکته لازم است که بسیاری از حملات مبتنی بر هوش مصنوعی به دلیل ماهیت پیچیده‌شان و زمینه مورد استفاده‌شان، در چندین تاکتیک همپوشانی دارند. در ادامه بر مبنای ترتیب مشخص‌شده در جدول ۲ شناسنامه حملات مشخص شده است.

جدول ۲: نگاشت حملات هوش مصنوعی تهاجمی با تاکتیک‌های ۱۴ گانه MITRE

تاکتیک	توضیحات تاکتیک	حملات مرتبط با هوش مصنوعی
جمع‌آوری اطلاعات (Recon-naissance)	جمع‌آوری اطلاعات برای برنامه‌ریزی عملیات آینده.	تحلیل ترافیک شبکه تشخیص آسیب‌پذیری جاسوسی پروفایل‌سازی هوشمند هدف‌محور تحلیل رفتاری برای یافتن راه‌های جدید سوءاستفاده پیش‌بینی نتایج حملات رمزشکنی کش‌کاوی مرد میانی
توسعه منابع (Resource Development)	ایجاد منابع برای حمایت از عملیات.	تولید خودکار دامنه مبهم‌سازی بدافزار تولید داده‌های مصنوعی تولید هوشمند نظرات جعلی حمله به CAPTCHA تولید خودکار اطلاعات نادرست سیبل
دسترسی اولیه (Initial Access)	کسب اولین پایگاه در سامانه یا شبکه هدف.	فیشینگ هدفمند حدس رمزعبور رمزشکنی بروت فورس رمزعبور حمله به CAPTCHA جعل جعل بیومتریک سرقت هویت تزریق اسکریپت‌های بین‌سایتی تزریق SQL جعل عمیق بازپخش سیبل مرد میانی تهدیدهای پایدار پیشرفته (APTs)[1]

1. Advanced Persistent Threat (APTs)

تاکتیک	توضیحات تاکتیک	حملات مرتبط با هوش مصنوعی
اجرا (Execution)	اجرای کد مخرب در سامانه هدف.	توزیع بدافزار بدافزارهای خودآموز تزریق SQL تزریق اسکریپت‌های بین‌سایتی بازپخش
پایداری (Persistence)	حفظ پایگاه در محیط هدف.	کلون‌سازی تهدیدهای پایدار پیشرفته (APTs)
ارتقاء دسترسی (Privilege Escalation)	کسب مجوزهای سطح بالاتر در سامانه.	تشخیص آسیب‌پذیری مهندسی معکوس کانال جانبی رمزشکنی کش‌کاوی
دور زدن دفاع (Defense Evasion)	اجتناب از تشخیص توسط مکانیزم‌های امنیتی.	بدافزارهای مبتنی بر هوش مصنوعی برای فرار از تشخیص دور زدن سامانه‌های تشخیص تهدیدات داخلی دور زدن فیلترهای ایمیل فرار از سامانه تشخیص نفوذ شبکه (IDS)[1] مبهم‌سازی بدافزار سازگارسازی حمله سیبل تهدیدهای پایدار پیشرفته (APTs) حمله به CAPTCHA
دسترسی به اعتبارنامه (Cre-dential Access)	سرقت اعتبارنامه‌ها برای دسترسی به سامانه‌ها.	سرقت رمز عبور رمزشکنی ثبت کلید ضمنی سرقت اعتبارنامه سرقت هویت بازپخش مرد میانی کش‌کاوی
کشف (Discovery)	کشف محیط هدف برای جمع‌آوری اطلاعات.	تحلیل ترافیک شبکه یافتن مسیر تحلیل رفتاری
حرکت جانبی (Lateral Movement)	حرکت در محیط هدف برای دسترسی به سامانه‌های اضافی.	حرکت جانبی یافتن مسیر
جمع‌آوری داده‌ها (Collection)	جمع‌آوری داده‌های مورد علاقه از هدف.	جاسوسی استخراج داده‌ها تغییر فیلم‌های نظارتی حمله حریم خصوصی

1. Intrusion Detection System (IDS)

تاکتیک	توضیحات تاکتیک	حملات مرتبط با هوش مصنوعی
فرمان و کنترل (Command and Control)	ارتباط با سامانه‌های مختل‌شده برای کنترل آن‌ها.	تولید دامنه خودکار هماهنگی حملات سیاه‌چاله
استخراج داده‌ها (Exfiltration)	سرقت داده‌ها از هدف.	استخراج داده‌ها دستکاری رکوردها رمزشکنی کش‌کاوی
تأثیر (Impact)	دستکاری، قطع یا تخریب سامانه‌ها یا داده‌های هدف.	حمله DDoS باج‌افزارهای مبتنی بر هوش مصنوعی دستکاری افکار عمومی دستکاری رکوردها سیاه‌چاله جمینگ تغییر فیلم‌های نظارتی تولید خودکار اطلاعات نادرست جعل عمیق بازپخش مرد میانی تهدیدهای پایدار پیشرفته (APTs) سیبل حمله به CAPTCHA

۲-۲ تحلیل ترافیک

حمله تحلیل ترافیک[1] مبتنی بر هوش مصنوعی نوعی پیشرفته از حملات سایبری است که در آن از هوش مصنوعی برای تحلیل و استنتاج اطلاعات حساس از الگوهای ترافیک شبکه استفاده می‌شود، حتی اگر محتوای ارتباطات رمزنگاری شده باشد. این نوع حمله بر روی متادیتا مانند اندازه بسته‌ها، زمان‌بندی، فرکانس و آدرس پروتکل اینترنت (IP)[2] مقصد تمرکز می‌کند تا فعالیت‌های کاربران، برنامه‌هایی که استفاده می‌کنند یا حتی محتوای خاصی که منتقل می‌شود را استنباط کند.

حملات تحلیل ترافیک خطرناک هستند زیرا می‌توانند با تمرکز بر ویژگی‌های قابل مشاهده، رمزنگاری را دور بزنند. هنگامی که این حملات با هوش مصنوعی ترکیب می‌شوند، دقیق‌تر، مقیاس‌پذیرتر و قادر به کشف بینش‌های عمیق‌تری از رفتار کاربران و عملیات سیستم می‌شوند.

۲-۲-۱ ویژگی‌های کلیدی

- **نظارت غیرفعال:**
  - مهاجم نیازی به مداخله فعال در شبکه ندارد و تنها ترافیک را به‌صورت غیرفعال مشاهده و تحلیل می‌کند.

1. Traffic Analysis Attack
2. Internet Protocol (IP)

- **سوءاستفاده از متادیتا:**
  - بر تحلیل متادیتا مانند اندازه بسته‌ها، زمان‌های بین ورود بسته‌ها، مدت زمان جریان‌ها و سایر ویژگی‌های غیرمرتبط با محتوا تمرکز می‌کند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی امکان تحلیل حجم بالایی از ترافیک را در زمان واقعی فراهم می‌کند و هدف قرار دادن کل شبکه‌ها یا اکوسیستم‌ها را ممکن می‌سازد.
- **پنهان‌کاری:**
  - از آنجا که این حمله تنها شامل مشاهده است و هیچ مداخله مستقیمی ندارد، تشخیص آن دشوار است.
- **انعطاف‌پذیری:**
  - مدل‌های هوش مصنوعی می‌توانند الگوهای جدید را یاد بگیرند و با آن‌ها سازگار شوند، دقت خود را در طول زمان بهبود بخشند و از اقدامات متقابل فرار کنند.

۲-۲-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - **یادگیری نظارت‌شده:** مدل‌ها را بر روی مجموعه‌داده‌های برچسب‌دار آموزش می‌دهد تا انواع مختلف ترافیک (مانند استریم ویدیو در مقابل مرور وب) را طبقه‌بندی کند.
  - **یادگیری بدون نظارت:** خوشه‌ها و ناهنجاری‌ها در الگوهای ترافیک را بدون برچسب‌گذاری قبلی شناسایی می‌کند.
  - **یادگیری تقویتی:** فرآیند تحلیل را با پاداش دادن به پیش‌بینی‌های دقیق و جریمه کردن پیش‌بینی‌های نادرست بهینه می‌کند.
- **یادگیری عمیق (DL)[1]:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی (CNN)[2] و شبکه‌های عصبی بازگشتی (RNNs)[3]، برای شناسایی الگوهای پیچیده در داده‌های ترتیبی مانند جریان‌های بسته استفاده می‌شوند.

1. Deep Learning (DL)
2. Convolutional Neural Networks (CNN)
3. Recurrent Neural Networks (RNNs)

- **پردازش زبان طبیعی (NLP)[1]:**
  - متادیتای متنی، مانند پرس‌وجوهای سیستم نام دامنه DNS[2] یا هدرهای HTTP، را تحلیل می‌کند تا زمینه‌های بیشتری درباره ترافیک استنباط کند.
- **شبکه‌های مولد تخاصمی (GANs)[3]:**
  - الگوهای ترافیک مصنوعی برای اهداف آموزشی یا آزمایش استحکام اقدامات دفاعی ایجاد می‌کند.
- **تحلیل سری‌های زمانی:**
  - از روش‌های آماری و یادگیری ماشین برای تحلیل وابستگی‌های زمانی در جریان‌های ترافیک استفاده می‌کند و به پیش‌بینی رفتارهای آینده کمک می‌کند.

۳-۲-۲ دارایی‌های هدف

- **شبکه‌های سازمانی:**
  - شبکه‌های شرکتی که شامل ارتباطات تجاری حساس، تراکنش‌های مالی یا مالکیت فکری هستند.
- **دستگاه‌های اینترنت اشیا (IoT)[4]:**
  - لوازم خانگی هوشمند، دستگاه‌های پوشیدنی و سنسورهای صنعتی که الگوهای ترافیک قابل پیش‌بینی ایجاد می‌کنند.
- **خدمات ابری:**
  - رابط برنامه‌نویسی کاربردی APIها[5]، خدمات ذخیره‌سازی و معماری‌های بدون سرور[6] که به پروتکل‌های ارتباطی رمزنگاری‌شده اما قابل پیش‌بینی متکی هستند.
- **ارائه‌دهندگان خدمات ارتباطی:**
  - شبکه‌های موبایل، ارائه‌دهندگان خدمات اینترنت[7] و ارائه‌دهندگان انتقال صدا از طریق پروتکل اینترنت (VoIP)[8] که زیرساخت آن‌ها حجم عظیمی از داده‌های کاربری را مدیریت می‌کند.

1. Natural Language Processing (NLP)
2. Domain Name System (DNS)
3. Generative Adversarial Network (GAN)
4. Internet of Things (IoT)
5. Application Programming Interface (API)
6. Serverless
7. ISPs
8. Voice over Internet Protocol(VOIP)

- **سازمان‌های دولتی:**
  - زیرساخت‌های حیاتی، ارتباطات نظامی و عملیات اطلاعاتی که به انتقال امن داده‌ها وابسته هستند.
- **سیستم‌های بهداشتی:**
  - دستگاه‌های پزشکی، پلتفرم‌های پزشکی از راه دور و سیستم‌های سوابق الکترونیک سلامت که داده‌های بیماران را منتقل می‌کنند.

۲-۲-۴ آسیب‌پذیری‌ها

- **عدم استفاده از ابهام‌سازی ترافیک:**
  - شبکه‌هایی که از تکنیک‌هایی مانند حاشیه‌سازی[1]، تصادفی‌سازی یا ترافیک ساختگی استفاده نمی‌کنند، تحلیل را آسان‌تر می‌کنند.
- **الگوهای قابل پیش‌بینی:**
  - فواصل منظم، اندازه‌های ثابت بسته‌ها یا مقاصد ثابت، پروفایل‌سازی ترافیک را آسان‌تر می‌کنند.
- **رمزنگاری ناکافی:**
  - در حالی که رمزنگاری از محتوا محافظت می‌کند، اغلب متادیتا را در معرض دید قرار می‌دهد و امکان استنتاج اطلاعات مفید را برای مهاجمان فراهم می‌کند.
- **کنترل‌های دسترسی ضعیف:**
  - دسترسی غیرمجاز به ابزارهای نظارتی یا تجهیزات شبکه، جمع‌آوری داده‌های ترافیک را برای مهاجمان ممکن می‌سازد.
- **خطای انسانی:**
  - تنظیمات نادرست فایروال‌ها، سیستم‌های ثبت گزارش‌ها[2] یا سیاست‌های امنیتی ممکن است به‌طور ناخواسته الگوهای ترافیک را در معرض دید قرار دهند.

۲-۲-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای نقشه‌برداری از شبکه هدف، شناسایی میزبان‌های فعال، الگوهای ارتباطی و دارایی‌های کلیدی استفاده می‌کند.

1. padding
2. Logs

- **جمع‌آوری داده:**
  - مهاجم ابزارهای شنود تقویت‌شده توسط یادگیری ماشین را برای ضبط فراداده از ترافیک شبکه، از جمله اندازه بسته‌ها، زمان‌بندی‌ها و جهت‌های جریان، مستقر می‌کند.
- **تحلیل و آموزش مدل:**
  - مهاجم از از الگوریتم‌های یادگیری ماشین برای تحلیل داده‌های جمع‌آوری‌شده استفاده می‌کند و الگوها را شناسایی کرده و فعالیت‌های کاربران یا وضعیت سیستم را استنباط می‌کند.
  - او مدل‌ها را آموزش داده تا برنامه‌ها، خدمات یا رفتارهای خاص را بر اساس داده‌های تاریخی تشخیص دهند.
- **استنتاج:**
  - مهاجم از الگوهای تحلیل‌شده برای استنتاج اطلاعات حساس، مانند نوع برنامه‌ای که استفاده می‌شود، ماهیت ارتباط یا هویت شرکت‌کنندگان، استفاده می‌کند.
- **سوءاستفاده:**
  - مهاجم از اطلاعات استنباط‌شده برای اهداف مخرب، مانند هدف قرار دادن دارایی‌های باارزش، انجام حملات مهندسی اجتماعی یا برنامه‌ریزی برای نفوذهای بیشتر، استفاده می‌کند.

۲-۲-۶ نشانه‌های احتمال نفوذ

- **الگوهای ترافیکی غیرعادی:**
  - ناهنجاری‌ها در اندازه بسته‌ها، زمان‌های بین ورود بسته‌ها یا مدت زمان جریان‌ها در مقایسه با رفتار پایه.
- **قرار گرفتن بیش از حد متادیتا در معرض دید:**
  - حجم زیادی از متادیتا که بدون ابهام‌سازی مناسب منتقل می‌شود.
- **تغییرات ناگهانی در حجم ترافیک:**
  - افزایش یا کاهش قابل توجه ترافیک که با رویدادها یا فعالیت‌های خاص همبستگی دارد.
- **ابزارهای نظارتی غیرمجاز:**
  - وجود ابزارهای ناشناخته یا غیرمجاز در شبکه که قادر به ضبط و تحلیل ترافیک هستند.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول کاربران، مانند الگوهای دسترسی غیرمعمول یا تعاملات

غیرمنتظره بین سیستم‌ها.

۲-۲-۷ تأثیرات

- **نقض حریم خصوصی:**
  - فعالیت‌ها، ترجیحات و عادات خصوصی کاربران ممکن است افشا شود و منجر به نقض حریم شخصی شود.
- **سرقت مالکیت فکری:**
  - ارتباطات تجاری حساس یا اطلاعات اختصاصی ممکن است استنباط و توسط رقبا مورد سوءاستفاده قرار گیرد.
- **ضررهای مالی:**
  - تراکنش‌های مالی یا جزئیات پرداخت ممکن است رهگیری و برای اهداف کلاهبرداری استفاده شود.
- **اختلال در عملیات:**

دانش فرآیندهای داخلی یا آسیب‌پذیری‌ها ممکن است حملات هدفمند به سیستم‌های حیاتی را ممکن سازد.

- **آسیب به اعتبار:**
  - نقض محرمانگی می‌تواند اعتماد مشتریان را از بین ببرد و به اعتبار سازمان آسیب برساند.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، می‌تواند منجر به جریمه‌ها و پیامدهای قانونی شود.

۲-۲-۸ راه‌های مقابله

- **مبهم‌سازی ترافیک:**
  - تکنیک‌هایی مانند padding، تصادفی‌سازی و ترافیک ساختگی را برای پنهان کردن الگوهای ارتباطی واقعی پیاده‌سازی کنید.
  - از رمزگذاری با نرخ بیت ثابت استفاده کنید تا اندازه بسته‌ها و فواصل یکنواخت باشند.
- **رمزنگاری انتها به انتها:**
  - تمام ارتباطات، از جمله متادیتا، را با استفاده از پروتکل‌های قوی مانند TLS 1.3 یا QUIC رمزنگاری کنید.

- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا میزان ترافیک قابل مشاهده در هر نقطه محدود شود.
- **کنترل دسترسی:**
  - دسترسی به ابزارهای نظارتی و تجهیزات شبکه را تنها به پرسنل مجاز محدود کنید.

- **تحلیل‌های رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای ایجاد خطوط پایه و تشخیص انحرافات نشان‌دهنده حملات تحلیل ترافیک استفاده کنید.
- **بررسی‌های منظم:**
  - ممیزی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **آموزش کارکنان:**
  - کارکنان را در مورد بهترین روش‌ها برای ایمن‌سازی شبکه‌ها و تشخیص علائم خطر آموزش دهید.
- **استفاده از فناوری‌های تقویت حریم خصوصی (PETs)[1]:**

از فناوری‌هایی مانند Tor، مسیریابی پیازی یا شبکه‌های ترکیبی برای ناشناس‌سازی ترافیک و محافظت در برابر تحلیل استفاده کنید.

۲-۳ تشخیص آسیب‌پذیری

حمله تشخیص آسیب‌پذیری[2] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای خودکارسازی و بهبود فرآیند شناسایی آسیب‌پذیری‌ها در نرم‌افزار، سخت‌افزار یا سیستم‌های شبکه است. برخلاف ابزارهای سنتی اسکن آسیب‌پذیری که به قواعد و امضاهای از پیش تعریف‌شده متکی هستند، حملات مبتنی بر هوش مصنوعی از یادگیری ماشین و یادگیری عمیق برای تحلیل الگوهای پیچیده، پیش‌بینی نقاط ضعف بالقوه و سازگاری با اقدامات امنیتی در حال تکامل استفاده می‌کنند. این امر باعث می‌شود مهاجمان بتوانند آسیب‌پذیری‌هایی را کشف و سوءاستفاده کنند که در غیر این صورت ممکن است نادیده گرفته شوند.

1. Privacy-Enhancing Technologies (PETs)
2. Vulnerability detection attack

این نوع حمله می‌تواند با کشف آسیب‌پذیری‌های روز صفر[1]، پیکربندی‌های نادرست یا نقص‌های ظریفی که تشخیص آن‌ها برای انسان‌ها دشوار است، از دفاع‌های امنیتی سنتی عبور کند.

۲-۳-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند زمان‌بر بررسی آسیب‌پذیری را خودکار می‌کند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **مقیاس‌پذیری:**
  - مدل‌های یادگیری ماشین می‌توانند مجموعه‌های بزرگ داده‌های پیکربندی سیستم، پایگاه‌های کد یا ترافیک شبکه را تحلیل کنند و به مهاجمان امکان می‌دهند تا چندین سیستم را به‌طور همزمان اسکن کنند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در نرم‌افزار، سخت‌افزار یا مکانیزم‌های دفاعی مانند مدیریت وصله‌ها یا سیستم‌های تشخیص نفوذ (IDS) سازگار شوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا آسیب‌پذیری‌های خاص را حتی در سیستم‌های پیچیده یا مبهم‌سازی‌شده شناسایی کنند.
- **پنهان‌کاری:**
  - با تقلید از فعالیت‌های اسکن قانونی یا استفاده از تکنیک‌های بررسی ظریف، حملات مبتنی بر هوش مصنوعی می‌توانند در مراحل شناسایی کشف نشوند.

۲-۳-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌های داده‌ای از آسیب‌پذیری‌های شناخته‌شده آموزش می‌دهد تا ویژگی‌های آن‌ها را یاد بگیرد و مسائل مشابه را در سیستم‌های جدید شناسایی کند.
  - یادگیری بدون نظارت: الگوها یا ناهنجاری‌ها را در داده‌های بدون برچسب شناسایی می‌کند

1. zero-day

و به کشف آسیب‌پذیری‌ها یا پیکربندی‌های نادرست پنهان کمک می‌کند.

- یادگیری تقویتی: فرآیند تشخیص آسیب‌پذیری را با پاداش دادن به کشف‌های موفقیت‌آمیز و جریمه کردن مثبت‌های کاذب بهینه می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در کد منبع، باینری‌ها یا ترافیک شبکه را تحلیل می‌کنند تا آسیب‌پذیری‌ها را تشخیص دهند.
  - شبکه‌های مولد تخاصمی: آسیب‌پذیری‌های مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا برای اهداف آموزشی یا آزمایش اقدامات دفاعی استفاده شوند.
- **پردازش زبان طبیعی:**
  - مستندات، نظرات یا پیام‌های خطا در پایگاه‌های کد را تحلیل می‌کند تا زمینه‌های اضافی درباره آسیب‌پذیری‌های بالقوه را استنباط کند.
- **تکنیک‌های فازی:**
  - هوش مصنوعی را با روش‌های فازی ترکیب می‌کند تا سیستم‌ها را به‌طور سیستماتیک برای رفتارهای غیرمنتظره یا خرابی‌ها آزمایش کند و نقص‌های منطقی زیرین را آشکار کند.
- **اجرای نمادین:**
  - از هوش مصنوعی برای شبیه‌سازی مسیرهای اجرای برنامه استفاده می‌کند و آسیب‌پذیری‌ها یا نقص‌های منطقی بالقوه را بدون نیاز به داده‌های زمان اجرا شناسایی می‌کند.

۲-۳-۳ دارایی‌های هدف

- **برنامه‌های نرم‌افزاری:**
  - برنامه‌های وب، برنامه‌های موبایل یا نرم‌افزارهای دسکتاپ که حاوی آسیب‌پذیری‌هایی مانند تزریق SQL، اسکریپت‌گذاری بین‌سایتی (XSS)[1] یا سرریز بافر هستند.
- **سیستم‌های شبکه:**
  - شبکه‌های سازمانی، پلتفرم‌های ابری یا اکوسیستم‌های اینترنت اشیا (IoT) که دارای سرویس‌های در معرض خطر، فایروال‌های نادرست پیکربندی‌شده یا پروتکل‌های قدیمی هستند.

1. Cross-Site Scripting (XSS)

- **دستگاه‌های سخت‌افزاری:**
  - سیستم‌های تعبیه‌شده، فرمور یا میکروکد کنترل‌کننده زیرساخت‌های حیاتی، دستگاه‌های پزشکی یا سیستم‌های خودرویی.
- **سیستم‌های رمزنگاری:**
  - طرح‌های رمزنگاری سفارشی یا پروتکل‌هایی که در نرم‌افزار یا سخت‌افزار پیاده‌سازی شده‌اند و ممکن است نقص‌های کشف‌نشده داشته باشند.
- **مؤلفه‌های زنجیره تأمین:**
  - کتابخانه‌ها، وابستگی‌ها یا مؤلفه‌های متن‌باز شخص ثالث که در سیستم‌های بزرگتر ادغام شده‌اند و ممکن است آسیب‌پذیری‌ها را معرفی کنند.

۴-۳-۲ آسیب‌پذیری‌ها

- **نرم‌افزارهای قدیمی:**
  - استفاده از کتابخانه‌ها، چارچوب‌ها یا سیستم‌های عامل قدیمی که حاوی آسیب‌پذیری‌های شناخته‌شده هستند.
- **پیکربندی ضعیف:**
  - فایروال‌ها، سرورها یا برنامه‌های نادرست پیکربندی‌شده که سرویس‌ها یا داده‌های حساس را در معرض خطر قرار می‌دهند.
- **تست ناکافی:**
  - عدم تست جامع برای آسیب‌پذیری‌ها در طول توسعه، سیستم‌ها را در معرض سوءاستفاده قرار می‌دهد.
- **خطای انسانی:**
  - اشتباهات توسعه‌دهندگان، مدیران یا اپراتورها ممکن است به‌طور ناخواسته آسیب‌پذیری‌هایی را در سیستم‌ها ایجاد کنند.
- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - آسیب‌پذیری‌های کشف‌نشده که هنوز راه‌حلی برای آن‌ها ارائه نشده یا به‌طور عمومی افشا نشده‌اند.

۲-۳-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره سیستم هدف، از جمله معماری، پیکربندی و رفتار آن استفاده می‌کند.
- **تحلیل:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های جمع‌آوری‌شده مستقر می‌کند و الگوها یا ناهنجاری‌هایی را که نشان‌دهنده آسیب‌پذیری‌ها هستند، شناسایی می‌کند.
- **سوءاستفاده:**
  - مهاجم از آسیب‌پذیری‌های شناسایی‌شده برای دسترسی غیرمجاز، سرقت داده یا اختلال در عملیات استفاده می‌کند.
- **پایداری:**
  - مهاجم درهای پشتی ایجاد می‌کند یا نقاط دسترسی را حفظ می‌کند تا امکان سوءاستفاده مکرر در طول زمان فراهم شود.
- **پوشاندن ردپا:**
  - مهاجم لاگ‌ها را تغییر می‌دهد، ردپاها را پاک می‌کند یا از تکنیک‌های دیگر برای اجتناب از تشخیص و حفظ پنهان‌کاری استفاده می‌کند.

۲-۳-۶ نشانه‌های احتمال نفوذ

- **فعالیت اسکن غیرمعمول:**
  - ناهنجاری‌ها در ترافیک شبکه، مانند بررسی‌های مکرر یا درخواست‌ها به نقاط پایانی غیرمعمول.
- **افزایش استفاده از منابع:**
  - مصرف بالاتر CPU یا حافظه به دلیل فعالیت‌های اسکن یا تحلیل خودکار.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول سیستم ناسازگار است، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به احراز هویت ناموفق، قفل شدن

حساب‌ها یا فعالیت‌های غیرمنتظره.

- **تغییرات غیرمنتظره:**
  - تغییرات ناگهانی در پیکربندی سیستم، مجوزهای فایل یا تنظیمات شبکه که توسط کاربران قانونی آغاز نشده‌اند.

۲-۳-۷ تأثیرات

- **دسترسی غیرمجاز:**
  - سوءاستفاده موفقیت‌آمیز از آسیب‌پذیری‌ها منجر به دسترسی غیرمجاز به حساب‌ها یا سیستم‌های حساس می‌شود.
- **نشت داده:**
  - سرقت داده‌های شخصی، مالی یا مالکیت فکری که منجر به خسارات مالی قابل توجه یا آسیب به اعتبار می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل به‌خطر افتادن اعتبارنامه‌ها یا سوءاستفاده از آسیب‌پذیری‌ها.
- **خسارات مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلال در خدمات که منجر به خسارات مالی می‌شود.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

۲-۳-۸ راه‌های مقابله

- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرم‌ورها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده را برطرف کنید.
- **پیکربندی ایمن:**
  - بهترین روش‌ها برای ایمن‌سازی سیستم‌ها، از جمله قواعد فایروال مناسب، کنترل‌های دسترسی و رمزنگاری را دنبال کنید.

- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت که نشان‌دهنده اسکن آسیب‌پذیری یا تلاش‌های سوءاستفاده هستند، استفاده کنید.
- **تست نفوذ:**
  - تست‌های نفوذ و ارزیابی‌های آسیب‌پذیری را به‌طور منظم انجام دهید تا نقاط ضعف را شناسایی و به‌طور پیش‌گیرانه برطرف کنید.
- **مدیریت وصله‌ها:**
  - فرآیندهای مدیریت وصله قوی را پیاده‌سازی کنید تا اطمینان حاصل شود که به‌روزرسانی‌های امنیتی به‌موقع اعمال می‌شوند.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد تشخیص علائم سوءاستفاده از آسیب‌پذیری‌ها و رعایت بهداشت امنیت سایبری آموزش دهید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا تأثیر یک حمله موفق را محدود کرده و جداسازی گره‌های به‌خطر افتاده را آسان‌تر کنید.
- **سیستم‌های تشخیص نفوذ (IDS):**
  - راه‌حل‌های پیشرفته IDS را مستقر کنید که قادر به تشخیص و پاسخ به حملات تشخیص آسیب‌پذیری مبتنی بر هوش مصنوعی در زمان واقعی باشند.

۲-۴ جاسوسی

حمله جاسوسی[1] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای خودکارسازی، بهبود و مقیاس‌پذیری تکنیک‌های سنتی جاسوسی است. این حملات فراتر از نظارت دستی یا جمع‌آوری داده‌ها هستند و از یادگیری ماشین و یادگیری عمیق برای تحلیل حجم عظیمی از داده‌ها، شناسایی الگوها و استنباط اطلاعات حساس درباره افراد، سازمان‌ها یا سیستم‌ها استفاده می‌کنند. با ترکیب الگوریتم‌های پیشرفته با ابزارهایی مانند تشخیص چهره، پردازش زبان طبیعی و تحلیل رفتاری، مهاجمان می‌توانند عملیات جاسوسی بسیار هدفمند و پنهان را انجام دهند.

این نوع حمله به‌ویژه در سناریوهای مربوط به جاسوسی شرکتی، عملیات سایبری حمایت‌شده توسط دولت یا نقض حریم خصوصی افراد خطرناک است. حملات جاسوسی مبتنی بر هوش مصنوعی

1. Spying attack

می‌توانند از دفاع‌های سنتی عبور کنند، با اقدامات ضد جاسوسی سازگار شوند و در مقیاس‌های بی‌سابقه‌ای عمل کنند.

۲-۴-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند جمع‌آوری داده، تحلیل و استنباط را خودکار می‌کند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **مقیاس‌پذیری:**
  - مدل‌های یادگیری ماشین می‌توانند مجموعه‌های بزرگ داده از منابع مختلف را به‌طور همزمان تحلیل کنند و به مهاجمان امکان می‌دهند تا به‌طور همزمان بر چندین هدف جاسوسی کنند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در اقدامات امنیتی، رفتار کاربر یا شرایط محیطی سازگار شوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا افراد، گروه‌ها یا سیستم‌های خاص را با دقت بالا شناسایی کنند.
- **پنهان‌کاری:**
  - با تقلید از فعالیت‌های قانونی یا ترکیب شدن با نویز پس‌زمینه، حملات جاسوسی مبتنی بر هوش مصنوعی می‌توانند برای مدت‌های طولانی کشف نشوند.

۲-۴-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌های داده‌ای از رفتارها، فعالیت‌ها یا ارتباطات شناخته‌شده آموزش می‌دهد تا الگوهای مشابه را در داده‌های جدید پیش‌بینی کند.
  - یادگیری بدون نظارت: ناهنجاری‌ها یا خوشه‌ها را در داده‌های بدون برچسب شناسایی می‌کند و به کشف روابط پنهان یا فعالیت‌های مشکوک کمک می‌کند.
  - یادگیری تقویتی: استراتژی‌های جاسوسی را با پاداش دادن به جمع‌آوری موفقیت‌آمیز اطلاعات و جریمه کردن تشخیص یا شکست بهینه می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در تصاویر، صدا، متن یا ترافیک شبکه را تحلیل می‌کنند تا اطلاعات حساس را تشخیص دهند.
  - شبکه‌های مولد تخاصمی: داده‌های مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا برای اهداف آموزشی یا آزمایش اقدامات دفاعی استفاده شوند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند ایمیل‌ها، لاگ‌های چت، پست‌های شبکه‌های اجتماعی یا اسناد را تحلیل می‌کند تا بینش‌ها، احساسات یا اطلاعات شخصی (PII)[1] را استخراج کند.
- **بینایی کامپیوتر:**
  - در تشخیص چهره، شناسایی اشیا یا ردیابی فعالیت استفاده می‌شود تا افراد یا محیط‌ها را به‌صورت بصری نظارت کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند و به مهاجمان امکان می‌دهد تا انحرافات را تشخیص دهند یا موجودیت‌های قانونی را جعل کنند.

۳-۴-۲ دارایی‌های هدف

- **افراد:**
  - کارمندان، مدیران یا سایر پرسنل کلیدی که ارتباطات، حرکات یا فعالیت‌های آن‌ها برای مهاجمان جذاب است.
- **سازمان‌ها:**
  - شرکت‌ها، سازمان‌های دولتی یا مؤسساتی که مالکیت فکری، اسرار تجاری یا داده‌های حساس را ذخیره می‌کنند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، سیستم‌های تأمین آب، شبکه‌های حمل‌ونقل یا تأسیسات بهداشتی که جاسوسی می‌تواند منجر به اختلالات قابل توجهی شود.

1. Personally Identifiable Information (PII)

- **دستگاه‌های اینترنت اشیا (IoT):**
  - لوازم خانگی هوشمند، دستگاه‌های پوشیدنی یا سنسورهای صنعتی که داده‌هایی درباره کاربران یا محیطها جمع‌آوری و انتقال می‌دهند.
- **سیستم‌های ارتباطی:**
  - سرورهای ایمیل، پلتفرم‌های پیام‌رسانی یا شبکه‌های ارتباطی که برای ارتباطات خصوصی یا محرمانه استفاده می‌شوند.

۲-۴-۴ آسیب‌پذیری‌ها

- **کنترل‌های ضعیف حریم خصوصی:**
  - عدم وجود مکانیزم‌های رمزنگاری قوی، کنترل دسترسی یا ناشناس‌سازی، داده‌های حساس را در معرض جاسوسی قرار می‌دهد.
- **نظارت ناکافی:**
  - روش‌های ضعیف ثبت لاگ یا عدم نظارت بلادرنگ، تشخیص و پاسخ به فعالیت‌های جاسوسی را دشوار می‌کند.
- **خطای انسانی:**
  - سیستم‌های نادرست پیکربندی‌شده، اعتبارنامه‌های اشتراکی یا آموزش ناکافی ممکن است به‌طور ناخواسته آسیب‌پذیری‌هایی را ایجاد کنند که از طریق جاسوسی قابل سوءاستفاده باشند.
- **نشت داده:**
  - مجموعه‌های داده نشت‌یافته، مواد ارزشمندی برای حملات جاسوسی مبتنی بر هوش مصنوعی فراهم می‌کنند و به مهاجمان امکان می‌دهند استراتژی‌های خود را بهبود بخشند.
- **عدم وجود اقدامات ضد جاسوسی:**
  - عدم وجود فناوری‌های ضد نظارت یا اقدامات دفاعی، جمع‌آوری غیرمجاز داده را تسهیل می‌کند.

۲-۴-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره هدف، از جمله حضور آنلاین، عادات ارتباطی یا مکان‌های فیزیکی استفاده می‌کند.

- **جمع‌آوری داده:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های عمومی، رهگیری ارتباطات یا سوءاستفاده از آسیب‌پذیری‌ها در سیستم‌ها مستقر می‌کند.
- **تحلیل:**
  - مهاجم الگوریتم‌های یادگیری عمیق را برای استخراج بینش‌های معنادار، شناسایی الگوها یا استنباط اطلاعات حساس از داده‌های جمع‌آوری‌شده اعمال می‌کند.
- **سوءاستفاده:**
  - مهاجم از اطلاعات جمع‌آوری‌شده برای اهداف مخرب، مانند باج‌گیری، جاسوسی شرکتی یا کسب مزیت استراتژیک استفاده می‌کند.
- **پایداری:**
  - مهاجم نظارت بلندمدت یا نقاط دسترسی را حفظ می‌کند تا اطمینان حاصل شود که جمع‌آوری داده به‌طور مداوم انجام می‌شود و با تغییر شرایط سازگار می‌شود.

۲-۴-۶ نشانه‌های احتمال نفوذ

- **الگوهای دسترسی غیرعادی به داده:**
  - ناهنجاری‌ها در پرس‌وجوهای پایگاه داده، دسترسی به فایل‌ها یا فراخوانی‌های واسط برنامه‌نویسی (API)[1] که با کاربران قانونی مرتبط نیستند.
- **افزایش ترافیک شبکه:**
  - ترافیک خروجی بیشتر از حد طبیعی به دلیل خروج داده‌های جمع‌آوری‌شده.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر ناسازگار است، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به دسترسی غیرمجاز، احراز هویت ناموفق یا فعالیت‌های غیرمنتظره.
- **شواهد فیزیکی:**
  - نشانه‌های دستکاری در دستگاه‌های اینترنت اشیا، دوربین‌ها یا سایر تجهیزات نظارتی.

1. Application Programming Interface(API)

۲-۴-۷ تأثیرات

- **نقض حریم خصوصی:**
  - جمع‌آوری و استفاده غیرمجاز از داده‌های شخصی که منجر به آسیب به اعتبار یا پیامدهای قانونی می‌شود.
- **سرقت مالکیت فکری:**
  - سرقت الگوریتم‌ها، طراحی‌ها یا منطق تجاری اختصاصی که منجر به خسارات مالی یا اختلال رقابتی می‌شود.
- **اختلال در عملیات:**
  - افشای آسیب‌پذیری‌ها یا نقاط ضعف در سیستم‌ها که امکان سوءاستفاده یا حملات بیشتر را فراهم می‌کند.
- **خسارات مالی:**
  - دسترسی غیرمجاز به سیستم‌های مالی یا سرقت دارایی‌های دیجیتال که باعث آسیب مالی می‌شود.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و اقدامات قانونی می‌شود.

۲-۴-۸ راه‌های مقابله

- **رمزنگاری:**
  - رمزنگاری انتها به انتها را برای تمام ارتباطات، ذخیره‌سازی داده و انتقال‌ها پیاده‌سازی کنید.
- **کنترل دسترسی:**
  - کنترل‌های دسترسی سختگیرانه، احراز هویت چندعاملی (MFA)[1] و اصل حداقل امتیاز را اعمال کنید تا دسترسی غیرمجاز محدود شود.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت که نشان‌دهنده تلاش‌های جاسوسی هستند، استفاده کنید.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی

1. Multi-Factor Authentication (MFA)

و برطرف کنید.

- **کاهش داده:**
  - فقط حداقل داده‌های لازم را جمع‌آوری و ذخیره کنید تا تأثیر احتمالی نفوذ کاهش یابد.
- **فناوری‌های ضد نظارت:**
  - ابزارهای ضد نظارت، مانند برنامه‌های پیام‌رسانی رمزنگاری‌شده، پروتکل‌های مرور ایمن یا نرم‌افزارهای ضد تشخیص چهره را مستقر کنید.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد تشخیص علائم جاسوسی و رعایت بهداشت امنیت سایبری آموزش دهید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش دسترسی غیرمجاز یا جمع‌آوری داده محدود شود.

۲-۵ پروفایل‌سازی هوشمند هدف‌محور

پروفایل‌سازی هوشمند هدف‌محور[1] با استفاده از هوش مصنوعی به فرآیند خودکارسازی و بهبود جمع‌آوری، تحلیل و تفسیر اطلاعات درباره اهداف بالقوه برای حملات سایبری اشاره دارد. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند پروفایل‌های بسیار دقیقی از افراد، سازمان‌ها یا سیستم‌ها ایجاد کنند که شامل شناسایی آسیب‌پذیری‌ها، ترجیحات، رفتارها و روابط می‌شود. این نوع پروفایل‌سازی به مهاجمان امکان می‌دهد حملات دقیق‌تر و مؤثرتری مانند فیشینگ هدفمند، مهندسی اجتماعی یا سرقت اعتبارنامه طراحی کنند.

پروفایل‌سازی هوشمند هدف‌محور به‌ویژه از داده‌های عمومی، روانشناسی انسان و روش‌های امنیتی ضعیف برای شناسایی اهداف باارزش و طراحی کمپین‌های مخرب استفاده می‌کند. خودکارسازی و مقیاس‌پذیری هوش مصنوعی این امکان را فراهم می‌کند که هزاران هدف به‌طور همزمان پروفایل‌سازی شوند، که این امر کارایی و تأثیر حملات سایبری را افزایش می‌دهد.

۲-۵-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند جمع‌آوری، تحلیل و تفسیر داده‌ها را خودکار می‌کند، که باعث کاهش

1. Intelligent Target Profiling

تلاش دستی و افزایش سرعت می‌شود.

- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا نقاط ضعف یا فرصت‌های خاص هر هدف را شناسایی کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی امکان پروفایل‌سازی سریع چندین هدف در صنایع یا جمعیت‌های مختلف را فراهم می‌کند.
- **انعطاف‌پذیری:**
  - پروفایل‌ها می‌توانند به‌طور مداوم بر اساس بازخوردهای لحظه‌ای یا تغییرات در محیط هدف به‌روزرسانی و بهبود یابند.
- **پنهان‌کاری:**
  - با ترکیب شدن در فعالیت‌های معمول جمع‌آوری داده‌ها یا استفاده از اطلاعات عمومی، هوش مصنوعی اطمینان می‌دهد که پروفایل‌سازی بدون تشخیص باقی بماند.

۲-۵-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های برچسب‌خورده از تلاش‌های موفق و ناموفق پروفایل‌سازی آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا خوشه‌ها را در داده‌های بدون ساختار شناسایی می‌کند تا روابط پنهان یا آسیب‌پذیری‌های بالقوه را آشکار کند.
  - یادگیری تقویتی: تکنیک‌های پروفایل‌سازی را با پاداش دادن به نتایج موفق و جریمه کردن شکست‌ها بهینه می‌کند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی از شبکه‌های اجتماعی، مقالات خبری، ایمیل‌ها و اسناد عمومی را تحلیل می‌کند تا بینش‌ها، احساسات یا اطلاعات شخصی (PII) را استخراج کند.
  - پیام‌ها یا محتوای واقع‌گرایانه و متناسب با زمینه ایجاد می‌کند تا در طول حملات با اهداف تعامل برقرار کند.

- **تحلیل شبکه‌های اجتماعی:**
  - از شبکه‌های عصبی گرافی (GNNs)[1] برای تحلیل ارتباطات، نقش‌ها و نفوذ در شبکه‌های اجتماعی استفاده می‌کند تا افراد یا سیستم‌های کلیدی را شناسایی کند.
- **داده‌کاوی:**
  - حجم عظیمی از داده‌ها را از پلتفرم‌های آنلاین، فروم‌ها و پایگاه‌داده‌های نفوذشده اسکن می‌کند تا پروفایل‌های جامعی از اهداف ایجاد کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار طبیعی کاربر یا سیستم استفاده می‌کند تا به مهاجمان امکان پیش‌بینی زمان و نحوه سوءاستفاده از آسیب‌پذیری‌ها را بدهد.

۳-۵-۲ دارایی‌های هدف

- **افراد:**
  - مدیران اجرایی، کارمندان یا سایر افراد باارزش که دسترسی به سیستم‌ها یا داده‌های حساس آن‌ها را به اهدافی جذاب تبدیل می‌کند.
- **سازمان‌ها:**
  - شرکت‌ها، سازمان‌های دولتی یا مؤسساتی که مالکیت فکری، اسرار تجاری یا داده‌های مشتریان ارزشمند را ذخیره می‌کنند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، مراکز بهداشتی، مؤسسات مالی یا سیستم‌های حمل‌ونقل که پروفایل‌سازی می‌تواند منجر به آسیب‌های جدی شود.
- **اجزای زنجیره تأمین:**
  - تأمین‌کنندگان، شرکا یا فروشندگان شخص ثالث که حساب‌های آن‌ها می‌تواند به عنوان نقطه ورود به اکوسیستم‌های بزرگ‌تر مورد سوءاستفاده قرار گیرد.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های اینترنت اشیا با منابع محدود که اغلب به عنوان بخشی از بات‌نت‌ها یا نقاط ورود برای حملات هدفمند استفاده می‌شوند.

1. Graph Neural Networks (GNNs)

۲-۵-۴ آسیب‌پذیری‌ها

- **داده‌های عمومی:**
  - شبکه‌های اجتماعی، سایت‌های شبکه‌سازی حرفه‌ای و اسناد عمومی منابع غنی از اطلاعات برای پروفایل‌سازی مبتنی بر هوش مصنوعی فراهم می‌کنند.
- **روانشناسی انسان:**
  - افراد بیشتر احتمال دارد که در دام حملات متناسب با علایق، عادات یا فعالیت‌های اخیر خود بیفتند.
- **کنترل‌های ضعیف حریم خصوصی:**
  - نبود رمزنگاری قوی، ناشناس‌سازی یا کنترل‌های دسترسی، داده‌های حساس را در معرض دسترسی غیرمجاز قرار می‌دهد.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های تشخیصی قدیمی یا نادرست، سازمان‌ها را در برابر تکنیک‌های پروفایل‌سازی تطبیقی آسیب‌پذیر می‌کند.
- **وابستگی بیش از حد به خودکارسازی:**
  - دفاع‌های سنتی ممکن است نتوانند تهدیدات ظریف یا در حال تحول ناشی از پروفایل‌سازی مبتنی بر هوش مصنوعی را تشخیص دهند.

۲-۵-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره هدف، از جمله حضور آنلاین، عادات ارتباطی و فعالیت‌های اخیر استفاده می‌کند.
- **تحلیل و آموزش:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های جمع‌آوری‌شده و آموزش بر روی الگوهای رفتار قانونی به‌کار گرفته تا استراتژی‌های حمله را بهبود بخشد.
- **ایجاد پروفایل:**
  - مهاجم پروفایل‌های دقیقی از اهداف ایجاد می‌کند که شامل آسیب‌پذیری‌ها، ترجیحات و بردارهای حمله بالقوه باشد.

- **برنامه‌ریزی حمله:**
  - مهاجم از پروفایل‌ها برای طراحی حملات شخصی‌سازی‌شده مانند ایمیل‌های فیشینگ هدفمند، صفحات ورود جعلی یا بدافزارهای سفارشی استفاده می‌کند.
- **اجرا:**
  - مهاجم حملات طراحی‌شده را به هدف تحویل می‌دهد و اطمینان حاصل می‌کند که با زمینه قربانی هماهنگ هستند و حداکثر تعامل را ایجاد می‌کنند.
- **پایداری:**
  - مهاجم پروفایل‌ها را به‌طور مداوم بر اساس بازخوردهای لحظه‌ای به‌روزرسانی و بهبود می‌بخشد تا اثربخشی بلندمدت حفظ شود.

۲-۵-۶ نشانه‌های احتمال نفوذ

- **الگوهای غیرمعمول دسترسی به داده:**
  - ناهنجاری‌ها در فعالیت‌های وب‌اسکرپینگ، فراخوانی‌های API یا پرس‌وجوهای پایگاه‌داده که افراد یا سازمان‌های خاص را هدف قرار می‌دهند.
- **فعالیت‌های افزایش‌یافته شناسایی:**
  - اسکن یا بررسی مشکوک خدمات شبکه، نقاط پایانی یا برنامه‌ها.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره یا تلاش‌های دسترسی غیرمجاز.
- **درخواست‌های غیرمنتظره داده:**
  - انحراف در درخواست‌های داده، مانند پرس‌وجوهای غیرمعمول به APIها یا پایگاه‌داده‌های عمومی که ممکن است نشان‌دهنده فعالیت پروفایل‌سازی باشد.

۲-۵-۷ تأثیرات

- **حملات هدفمند:**
  - حملات بسیار شخصی‌سازی‌شده احتمال موفقیت را افزایش می‌دهند و منجر به نقض

داده‌ها، سرقت اعتبارنامه یا کلاهبرداری مالی می‌شوند.

- **اختلال عملیاتی:**
  - سیستم‌ها یا حساب‌های نفوذشده خدمات حیاتی کسب‌وکار را مختل می‌کنند یا باعث آسیب‌های اعتباری می‌شوند.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلالات خدمات منجر به آسیب‌های مالی می‌شوند.
- **آسیب به شهرت:**
  - قربانیان پروفایل‌سازی مبتنی بر هوش مصنوعی ممکن است به دلیل داده‌های افشاشده یا دستکاری‌شده آسیب اعتباری ببینند.
- **آسیب‌پذیری‌های سیستم:**
  - پروفایل‌سازی نقاط ضعف سیستم‌ها را آشکار می‌کند و به مهاجمان امکان می‌دهد از آن‌ها برای نفوذ بیشتر یا حرکت جانبی استفاده کنند.

۸-۵-۲ راه‌های مقابله

- **کاهش داده:**
  - میزان داده‌های شخصی یا سازمانی که به‌صورت عمومی به‌اشتراک گذاشته می‌شود را محدود کنید تا مواد قابل استفاده برای پروفایل‌سازی کاهش یابد.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی برای نظارت بر فعالیت‌های پروفایل‌سازی و بردارهای حمله نوظهور استفاده کنید.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرمعمول فعالیت که نشان‌دهنده پروفایل‌سازی یا شناسایی هستند استفاده کنید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرم‌ورها و پیکربندی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده توسط پروفایل‌سازی مبتنی بر هوش مصنوعی اصلاح شوند.

- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش دسترسی غیرمجاز یا جمع‌آوری داده محدود شود.
- **رمزنگاری قوی:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حال انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده کاهش یابد.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد تشخیص علائم پروفایل‌سازی، مانند درخواست‌های مشکوک برای اطلاعات یا رفتار غیرمعمول سیستم آموزش دهید.
- **نظارت پیش‌گیرانه:**
  - راه‌حل‌های نظارت مداوم را پیاده‌سازی کنید که قادر به تشخیص و پاسخ به تلاش‌های پروفایل‌سازی در زمان واقعی باشند.
- **کنترل‌های دسترسی:**
  - کنترل‌های دسترسی سخت‌گیرانه، احراز هویت چندعاملی و اصل حداقل امتیاز را اعمال کنید تا دسترسی غیرمجاز محدود شود.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حملات مبتنی بر پروفایل‌سازی را تشخیص، مهار و کاهش دهید.

۲-۶ تحلیل رفتاری برای یافتن راه‌های جدید سوءاستفاده

تحلیل رفتاری[1] در امنیت سایبری نه تنها یک ابزار دفاعی است، بلکه می‌تواند توسط مهاجمان به عنوان یک سلاح برای کشف راه‌های جدید سوءاستفاده از سیستم‌ها، کاربران یا شبکه‌ها مورد استفاده قرار گیرد. با بهره‌گیری از یادگیری ماشین، یادگیری عمیق و هوش مصنوعی، مهاجمان می‌توانند الگوهای رفتار کاربر و سیستم را تحلیل کنند تا آسیب‌پذیری‌ها را شناسایی کنند، اقدامات دفاعی را پیش‌بینی کنند و حملات پیچیده‌ای طراحی کنند که از تشخیص فرار می‌کنند. این شکل از تحلیل رفتاری به مهاجمان امکان می‌دهد از تکنیک‌های سنتی سوءاستفاده فراتر رفته و بر روش‌های ظریف و مبتنی بر زمینه تمرکز کنند که به ظرافت‌های رفتار انسان و ماشین حمله می‌کنند.

وقتی تحلیل رفتاری به‌طور مخرب استفاده شود، به یک ابزار قدرتمند شناسایی و برنامه‌ریزی

1. Behavior analysis

حمله تبدیل می‌شود که به مهاجمان امکان می‌دهد استراتژی‌های خود را به‌طور پویا تطبیق دهند و تأثیر عملیات خود را به حداکثر برسانند.

۲-۶-۱ ویژگی‌های کلیدی

- **شناسایی پیش‌گیرانه:**
  - رفتار کاربر و سیستم را تحلیل می‌کند تا نقاط ضعف بالقوه را قبل از اینکه توسط مدافعان مورد سوءاستفاده قرار گیرند، آشکار کند.
- **خودکارسازی:**
  - هوش مصنوعی فرآیند جمع‌آوری، پردازش و تفسیر حجم عظیمی از داده‌های رفتاری را خودکار می‌کند، که باعث کاهش تلاش دستی و افزایش کارایی می‌شود.
- **انعطاف‌پذیری:**
  - مدل‌های رفتاری با گذشت زمان تکامل می‌یابند و با تغییرات در رفتار کاربر، پیکربندی سیستم‌ها یا شرایط محیطی سازگار می‌شوند.
- **دقت:**
  - الگوریتم‌های پیشرفته امکان شناسایی دقیق رفتارهای قابل سوءاستفاده را فراهم می‌کنند و در عین حال منفی‌های کاذب را به حداقل می‌رسانند.
- **یادگیری مستمر:**
  - بازخوردهای لحظه‌ای از تعاملات با اهداف، استراتژی‌های حمله را بهبود می‌بخشد و موفقیت پایدار را تضمین می‌کند.

۲-۶-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های برچسب‌خورده از تلاش‌های موفق و ناموفق سوءاستفاده آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در داده‌های بدون برچسب شناسایی می‌کند تا آسیب‌پذیری‌های پنهان یا فرصت‌ها را آشکار کند.
  - یادگیری تقویتی: تاکتیک‌های سوءاستفاده را با پاداش دادن به نتایج موفق و جریمه کردن شکست‌ها بهینه می‌کند، که امکان بهبود مستمر را فراهم می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و شبکه ترنسفورمرها، وابستگی‌های زمانی در تعاملات کاربر یا سیستم را تحلیل می‌کنند تا الگوهای ظریف را تشخیص دهند یا اقدامات آینده را پیش‌بینی کنند.
  - شبکه‌های مولد تخاصمی: داده‌های رفتاری مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های فرار را آزمایش و بهبود بخشند یا بردارهای حمله جدید را شبیه‌سازی کنند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند ایمیل‌ها، لاگ‌های چت یا مستندات را تحلیل می‌کند تا اطلاعات بیشتری درباره آسیب‌پذیری‌های بالقوه یا اقدامات دفاعی استنباط کند.
- **تحلیل شبکه‌های اجتماعی:**
  - از شبکه‌های عصبی گرافی برای تحلیل روابط بین کاربران، دستگاه‌ها یا سیستم‌ها استفاده می‌کند تا اهداف باارزش یا ارتباطات مورد اعتماد را که می‌توانند مورد سوءاستفاده قرار گیرند، شناسایی کند.
- **تقلید رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار طبیعی کاربر یا سیستم استفاده می‌کند تا مهاجمان بتوانند اقداماتی را طراحی کنند که با الگوهای مورد انتظار هماهنگ باشند و در عین حال اهداف خود را پیش ببرند.

۳-۶-۲ دارایی‌های هدف

- **رفتار کاربر:**
  - عادات، ترجیحات و روال‌های فردی را نظارت می‌کند تا کمپین‌های فیشینگ شخصی‌سازی‌شده یا حملات مهندسی اجتماعی طراحی کند.
- **رفتار سیستم:**
  - نحوه تعامل سیستم‌ها با کاربران، دستگاه‌ها یا برنامه‌ها را تحلیل می‌کند تا پیکربندی‌های نادرست، مکانیزم‌های احراز هویت ضعیف یا آسیب‌پذیری‌های نادیده گرفته‌شده را آشکار کند.
- **ترافیک شبکه:**
  - الگوهای ارتباطی درون شبکه‌ها را بررسی می‌کند تا نقاط ورودی بالقوه، مسیرهای حرکت

جانبی یا مسیرهای استخراج داده را شناسایی کند.

- **زیرساخت‌های حیاتی:**
  - گردش‌های عملیاتی در شبکه‌های برق، مراکز بهداشتی، مؤسسات مالی یا شبکه‌های حمل‌ونقل را مطالعه می‌کند تا شکاف‌های قابل سوءاستفاده را پیدا کند.
- **سیستم‌های زنجیره تأمین:**
  - تأمین‌کنندگان، شرکا یا فروشندگان شخص ثالث را بررسی می‌کند تا حلقه‌های ضعیفی را که می‌توانند به عنوان نقاط ورودی برای اکوسیستم‌های بزرگتر مورد استفاده قرار گیرند، شناسایی کنند.

۲-۶-۴ آسیب‌پذیری‌ها

- **روانشناسی انسان:**
  - از بینش‌های مربوط به رفتار کاربر استفاده می‌کند تا پیام‌های فیشینگ بسیار متقاعدکننده ایجاد کند، افراد مورد اعتماد را جعل کند یا فرآیندهای تصمیم‌گیری را دستکاری کند.
- **مکانیزم‌های احراز هویت ضعیف:**
  - الگوهای قابل پیش‌بینی ورود به سیستم، اعتبارنامه‌های مشترک یا پیاده‌سازی‌های ناکافی احراز هویت چندعاملی را شناسایی می‌کند.
- **نظارت ناکافی:**
  - از روش‌های ضعیف ثبت لاگ یا نبود نظارت لحظه‌ای سوءاستفاده می‌کند تا در طول مراحل شناسایی و سوءاستفاده بدون تشخیص باقی بماند.
- **دفاع‌های قدیمی:**
  - سیستم‌های قدیمی یا نرم‌افزارهای منسوخ را هدف قرار می‌دهد که در آن‌ها ابزارهای تحلیل رفتاری ممکن است پیاده‌سازی یا به‌درستی پیکربندی نشده باشند.
- **آسیب‌پذیری‌های روز صفر:**
  - از داده‌های رفتاری برای پیش‌بینی یا کشف نقص‌های ناشناخته در سیستم‌ها یا فرآیندها استفاده می‌کند که می‌توانند برای اهداف مخرب مورد سوءاستفاده قرار گیرند.

۲-۶-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره کاربران، سیستم‌ها

یا شبکه‌های هدف، از جمله رفتارهای معمول، الگوهای دسترسی و سبک‌های ارتباطی آن‌ها استفاده می‌کند.

- **تحلیل و مدل‌سازی:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های جمع‌آوری‌شده مستقر کرده و خطوط پایه رفتار طبیعی را ایجاد و انحرافاتی را که نشان‌دهنده نقاط ضعف قابل سوءاستفاده هستند، شناسایی می‌کند.
- **برنامه‌ریزی سوءاستفاده:**
  - مهاجم از بینش‌های حاصل از تحلیل رفتاری برای طراحی حملات سفارشی‌سازی‌شده برای اهداف خاص، مانند فیشینگ هدفمند، حملات سرقت اعتبارنامه‌ها یا حرکت جانبی استفاده می‌کند.
- **اجرا:**
  - مهاجم حملات برنامه‌ریزی‌شده را اجرا کرده و اطمینان حاصل می‌کند که با زمینه قربانی هماهنگ هستند و خطر تشخیص را به حداقل می‌رسانند.
- **پایداری:**
  - مهاجم کمپین را به‌طور مداوم بر اساس بازخوردهای لحظه‌ای بهبود داده تا دسترسی بلندمدت یا تأثیر بر محیط‌های هدف حفظ شود.

۲-۶-۶ نشانه‌های احتمال نفوذ

- **الگوهای فعالیت غیرمعمول:**
  - ناهنجاری‌ها در دسترسی به فایل‌ها، پرس‌وجوهای پایگاه‌داده یا فراخوانی‌های API که با عملیات قانونی مرتبط نیستند.
- **افزایش فعالیت‌های شناسایی:**
  - اسکن یا بررسی مشکوک خدمات شبکه، نقاط پایانی یا برنامه‌ها.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره یا تلاش‌های دسترسی غیرمجاز.

- **درخواست‌های داده غیرمنتظره:**
  - انحراف در درخواست‌های داده، مانند پرس‌وجوهای غیرمعمول به APIها یا پایگاه‌داده‌های عمومی، که ممکن است نشان‌دهنده آماده‌سازی برای سوءاستفاده باشد.

۲-۶-۷ تأثیرات

- **حملات هدفمند:**
  - حملات بسیار شخصی‌سازی‌شده احتمال موفقیت را افزایش می‌دهند و منجر به نقض داده‌ها، سرقت اعتبارنامه یا کلاهبرداری مالی می‌شوند.
- **اختلال عملیاتی:**
  - سیستم‌ها یا حساب‌های نفوذشده خدمات حیاتی کسب‌وکار را مختل می‌کنند یا باعث آسیب‌های اعتباری می‌شوند.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلالات خدمات منجر به آسیب‌های مالی می‌شوند.
- **آسیب به شهرت:**
  - قربانیان سوءاستفاده رفتاری مبتنی بر هوش مصنوعی ممکن است به دلیل داده‌های افشاشده یا دستکاری‌شده آسیب اعتباری ببینند.
- **آسیب‌پذیری‌های سیستم:**
  - تحلیل رفتاری نقاط ضعف سیستم‌ها را آشکار می‌کند و به مهاجمان امکان می‌دهد از آن‌ها برای نفوذ بیشتر یا حرکت جانبی استفاده کنند.

۲-۶-۸ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به تشخیص رفتارهای غیرعادی نشان‌دهنده تلاش‌های سوءاستفاده هستند، حتی اگر اقدامات قانونی به نظر برسند.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرم‌ورها و پیکربندی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده از طریق تحلیل رفتاری اصلاح شوند.

- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای نظارت بر فعالیت کاربر و سیستم برای انحرافاتی که ممکن است نشان‌دهنده قصد مخرب باشد استفاده کنید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش دسترسی غیرمجاز یا فعالیت‌های مخرب محدود شود.
- **رمزنگاری قوی:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حال انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده کاهش یابد.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد تشخیص علائم سوءاستفاده، مانند تلاش‌های فیشینگ یا رفتار غیرمعمول سیستم آموزش دهید.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا درباره تکنیک‌های نوظهور سوءاستفاده و تهدیدات تطبیقی به‌روز بمانید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حملات مبتنی بر رفتار را تشخیص، مهار و کاهش دهید.
- **دفاع پیش‌گیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های در حال تحول سوءاستفاده رفتاری استفاده کنید.
- **تحلیل رفتار کاربر و موجودیت (UEBA):**
  - راه‌حل‌های UEBA[1] را پیاده‌سازی کنید که هوش مصنوعی را با مدل‌سازی رفتاری ترکیب می‌کنند تا تهدیدات داخلی، APTها یا سایر حملات پیشرفته را تشخیص دهند.

۲-۷ پیش‌بینی نتایج حملات

پیش‌بینی نتایج حملات[2] شامل استفاده از هوش مصنوعی برای پیش‌بینی موفقیت یا شکست یک

1. User and Entity Behavior Analytics (UBEA)
2. Attacks Outcome prediction

حمله سایبری است که به مهاجمان امکان می‌دهد تا استراتژی‌های خود را بهبود بخشند و تأثیر خود را به حداکثر برسانند. با به‌کارگیری یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند حجم عظیمی از داده‌های مربوط به آسیب‌پذیری‌های سیستم، رفتار کاربران، پیکربندی شبکه و اقدامات دفاعی را تحلیل کنند تا پیش‌بینی کنند که یک حمله چگونه پیش خواهد رفت. این نوع پیش‌بینی به مهاجمان اجازه می‌دهد تا دفاع‌ها را پیش‌بینی کنند، تاکتیک‌ها را در زمان واقعی تطبیق دهند و احتمال دستیابی به اهداف خود را افزایش دهند.

پیش‌بینی نتایج حملات مبتنی بر هوش مصنوعی تحلیل‌های پیشرفته را با اتوماسیون ترکیب می‌کند و به مهاجمان این توانایی را می‌دهد که کمپین‌های بسیار هدفمند و کارآمدی را اجرا کنند که از اقدامات امنیتی سنتی عبور می‌کنند.

۲-۷-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند جمع‌آوری داده‌ها، تحلیل سناریوها و پیش‌بینی نتایج را خودکار می‌کند، تلاش دستی را کاهش می‌دهد و سرعت را افزایش می‌دهد.
- **انعطاف‌پذیری:**
  - پیش‌بینی‌ها به‌طور پویا بر اساس بازخوردهای زمان واقعی تکامل می‌یابند و با تغییرات در دفاع‌های سیستم یا شرایط محیطی تطبیق می‌یابند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا آسیب‌پذیری‌ها یا شکاف‌های دفاعی خاص را شناسایی کنند تا حداکثر تأثیر را داشته باشند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا نتایج را در چندین سیستم یا شبکه به‌طور همزمان پیش‌بینی کنند و آسیب‌های بالقوه را افزایش دهند.
- **پنهان‌کاری:**
  - با پیش‌بینی مکانیزم‌های تشخیص، هوش مصنوعی اطمینان می‌دهد که حملات در حین پیشروی به سمت اهداف خود کشف نشوند.

۲-۷-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های حملات موفق و ناموفق آموزش می‌دهد

تا الگوها را یاد بگیرد و پیش‌بینی‌ها را بهبود بخشد.

- یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در رفتار سیستم را شناسایی می‌کند تا نقاط ضعف یا فرصت‌های پنهان را کشف کند.
- یادگیری تقویتی: استراتژی‌های حمله را با پاداش دادن به پیش‌بینی‌های موفق و جریمه کردن شکست‌ها بهینه می‌کند و امکان بهبود مستمر را فراهم می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و شبکه ترنسفورمرها، وابستگی‌های زمانی در تعاملات سیستم را تحلیل می‌کنند تا دقت را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: سناریوهای حمله مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های پیش‌بینی را آزمایش و بهبود بخشند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا به مهاجمان امکان پیش‌بینی پاسخ‌های دفاعی یا زمان‌بندی بهینه حمله را بدهد.
- **داده‌کاوی:**
  - حجم زیادی از داده‌های سیستم‌های به‌خطرافتاده، نفوذها یا منابع عمومی را تحلیل می‌کند تا مدل‌های پیش‌بینی را تغذیه کند.
- **شبیه‌سازی‌ها و سناریوها:**
  - از هوش مصنوعی برای شبیه‌سازی سناریوهای حمله و ارزیابی نتایج احتمالی قبل از اجرا استفاده می‌کند.

۳-۷-۲ دارایی‌های هدف

- **شبکه‌های سازمانی:**
  - سیستم‌های شرکتی که پیش‌بینی نتایج می‌تواند حملات چندوجهی مانند باج‌افزار یا APTها را هدایت کند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تأسیسات بهداشتی، مؤسسات مالی یا سیستم‌های حمل‌ونقل که پیش‌بینی‌های دقیق می‌تواند آسیب‌های قابل توجهی ایجاد کند.
- **آژانس‌های دولتی:**
  - اهداف باارزش بالا که اطلاعات طبقه‌بندی‌شده را ذخیره می‌کنند یا سیستم‌های امنیت ملی

را کنترل می‌کنند.

- **اکوسیستم‌های اینترنت اشیا (IoT):**
  - دستگاه‌هایی با منابع محدود که اغلب به‌عنوان نقطه ورود برای کمپین‌های بزرگتر استفاده می‌شوند و پیش‌بینی‌ها به اولویت‌بندی اهداف کمک می‌کنند.
- **سیستم‌های زنجیره تأمین:**
  - فروشندگان، شرکا یا تأمین‌کنندگان شخص ثالث که حساب‌های به‌خطرافتاده آن‌ها به‌عنوان پل‌هایی برای سوءاستفاده گسترده‌تر استفاده می‌شوند.

۲-۷-۴ آسیب‌پذیری‌ها

- **اقدامات دفاعی ضعیف:**
  - عدم وجود ابزارهای نظارتی قوی یا سیستم‌های تشخیص ناهنجاری، سازمان‌ها را در معرض حملات پیش‌بینی‌شده قرار می‌دهد.
- **خطای انسانی:**
  - کارمندانی که در دام ایمیل‌های فیشینگ می‌افتند، از رمزهای عبور ضعیف استفاده می‌کنند یا سیستم‌ها را به‌اشتباه پیکربندی می‌کنند، به‌طور ناخواسته آسیب‌پذیری‌هایی را در معرض سوءاستفاده از طریق پیش‌بینی قرار می‌دهند.
- **دفاع‌های قدیمی:**
  - سیستم‌های قدیمی یا نرم‌افزارهای منسوخ، تهدیدهای تطبیقی که توسط پیش‌بینی‌های مبتنی بر هوش مصنوعی فعال می‌شوند را در نظر نمی‌گیرند.
- **اتکای بیش از حد به اتوماسیون:**
  - دفاع‌های سنتی که به قوانین یا امضاهای ثابت متکی هستند، ممکن است تهدیدهای ظریف پیش‌بینی‌شده توسط هوش مصنوعی را از دست بدهند.
- **اطلاعات تهدید ناکافی:**
  - عدم وجود اطلاعات تهدید پیش‌گیرانه، سازمان‌ها را برای استراتژی‌های حمله در حال تحول که توسط پیش‌بینی‌ها هدایت می‌شوند، آماده نمی‌کند.

۲-۷-۵ سناریوی تهدید

- **جمع‌آوری داده‌ها:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره سیستم‌های

هدف، از جمله معماری، پیکربندی‌ها و رفتارهای معمول آن‌ها استفاده می‌کند.

- **تحلیل و آموزش:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های جمع‌آوری‌شده و آموزش روی الگوهای فعالیت قانونی در مقابل نتایج حمله مستقر می‌کند.
- **پیش‌بینی نتایج:**
  - مهاجم از هوش مصنوعی برای پیش‌بینی نرخ موفقیت انواع بردارهای حمله، مانند فیشینگ، استقرار بدافزار یا افزایش امتیاز استفاده می‌کند.
- **برنامه‌ریزی حمله:**
  - مهاجم استراتژی‌های حمله را بر اساس پیش‌بینی‌ها بهبود می‌بخشد و اطمینان حاصل می‌کند که اقدامات با آسیب‌پذیری‌های پیش‌بینی‌شده و پاسخ‌های دفاعی همسو هستند.
- **اجرا:**
  - مهاجم حمله برنامه‌ریزی‌شده را اجرا کرده و تاکتیک‌ها را بر اساس پیش‌بینی‌ها و بازخوردهای جاری به‌طور پویا تنظیم می‌کند.
- **تداوم:**
  - مهاجم پیش‌بینی‌ها و استراتژی‌ها را به‌طور مداوم به‌روزرسانی می‌کند تا دسترسی بلندمدت یا تأثیر بر محیط‌های هدف حفظ شود.

۲-۷-۶ نشانه‌های احتمال نفوذ

- **فعالیت غیرمعمول شناسایی:**
  - ناهنجاری‌ها در رفتارهای اسکن یا کاوش، نشان‌دهنده تلاش‌ها برای جمع‌آوری داده‌ها برای مدل‌سازی پیش‌بینی.
- **الگوهای افزایش دسترسی به داده‌ها:**
  - پرس‌وجوهای بیشتر از حد معمول به پایگاه‌های داده، APIها یا لاگ‌ها، که ممکن است نشان‌دهنده آماده‌سازی برای حملات پیش‌بینی‌شده باشد.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول کاربر یا سیستم، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.

- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.
- **الگوهای ترافیک غیرمنتظره:**
  - انحراف در ترافیک شبکه، مانند اتصالات به آدرس‌های پروتکل اینترنت IP یا دامنه‌های ناشناخته، که ممکن است نشان‌دهنده ارتباطات فرمان و کنترل (C2)[1] یا خروج داده باشد.

۲-۷-۷ تأثیرات

- **اختلال عملیاتی:**
  - پیش‌بینی‌های دقیق به مهاجمان امکان می‌دهد تا خدمات یا سیستم‌های حیاتی کسب‌وکار را با حداقل تلاش مختل کنند.
- **نشت داده‌ها:**
  - هدف‌گیری پیش‌بینی‌شده احتمال سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری را افزایش می‌دهد.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، کلاهبرداری یا اختلال‌های خدمات ناشی از حملات دقیق اجراشده منجر به آسیب‌های مالی می‌شود.
- **آسیب به اعتبار:**
  - قربانیان حملات پیش‌بینی‌شده به دلیل داده‌های افشا یا دستکاری‌شده، آسیب‌های اعتباری می‌بینند.
- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا فرم‌ورهای حیاتی، که سیستم‌ها را غیرقابل اعتماد یا غیرقابل استفاده می‌کند.

۲-۷-۸ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به شناسایی ناهنجاری‌های ظریف در ترافیک شبکه، رفتار کاربر یا لاگ‌های سیستم هستند که نشان‌دهنده حملات

1. Command and Control (C2)

پیش‌بینی‌شده است.

- **اطلاعات تهدید پیش‌گیرانه:**
  - از پلتفرم‌های اطلاعات تهدید مبتنی بر هوش مصنوعی استفاده کنید تا درباره تکنیک‌های پیش‌بینی‌شده و تهدیدهای تطبیقی در حال ظهور مطلع شوید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرم‌ورها و پیکربندی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده که در طول برنامه‌ریزی پیش‌بینی مورد سوءاستفاده قرار می‌گیرند، اصلاح شوند.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی استفاده کنید تا انحرافات در فعالیت کاربر یا سیستم که ممکن است نشان‌دهنده قصد مخرب باشد را نظارت کنید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش دسترسی غیرمجاز یا فعالیت‌های مخرب هدایت‌شده توسط پیش‌بینی‌ها محدود شود.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد تشخیص نشانه‌های خطر آموزش دهید و فرهنگ هوشیاری را تقویت کنید.
- **رمزنگاری پیشرفته:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حالت انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده را به حداقل برسانید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و اجرا کنید تا به‌سرعت حملات پیش‌بینی‌شده را تشخیص، مهار و کاهش دهید.
- **معماری اعتماد صفر:**
  - رویکرد اعتماد صفر را برای امنیت شبکه اتخاذ کنید و هر تلاش دسترسی را بدون توجه به منشأ آن تأیید کنید.
- **آزمایش دفاع شبیه‌سازی‌شده:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های حمله پیش‌بینی‌شده در حال تحول استفاده کنید.

۲-۸ رمزشکنی

رمزشکنی[1] مطالعه روش‌هایی برای به‌دست‌آوردن معنای اطلاعات رمزنگاری‌شده بدون دسترسی به کلید مخفی است. در حمله رمزشکنی مبتنی بر هوش مصنوعی، از تکنیک‌های هوش مصنوعی برای بهبود و خودکارسازی فرآیند شکستن الگوریتم‌ها یا پروتکل‌های رمزنگاری استفاده می‌شود. این حملات از ضعف‌های موجود در طرح‌های رمزنگاری، اشکالات پیاده‌سازی یا آسیب‌پذیری‌های کانال جانبی برای بازیابی متن‌های آشکار، کلیدها یا سایر داده‌های حساس سوءاستفاده می‌کنند.

با پیشرفت‌های یادگیری ماشین و یادگیری عمیق، رمزشکنی مبتنی بر هوش مصنوعی قدرتمندتر شده و قادر به تحلیل حجم عظیمی از متن‌های رمزنگاری‌شده، شناسایی الگوها و پیش‌بینی ضعف‌های احتمالی است که تشخیص دستی آن‌ها برای انسان‌ها غیرعملی است. این موضوع، رمزشکنی مبتنی بر هوش مصنوعی را به یک تهدید جدی برای سیستم‌های رمزنگاری مدرن تبدیل کرده است.

۲-۸-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند زمان‌بر رمزشکنی را خودکار می‌کند و به مهاجمان امکان می‌دهد تا فرضیه‌ها را آزمایش کرده و استراتژی‌ها را در مقیاس بزرگ بهبود بخشند.
- **مقیاس‌پذیری:**
  - مدل‌های یادگیری ماشین می‌توانند مجموعه‌داده‌های بزرگ متن‌های رمزنگاری‌شده را تحلیل کنند و شناسایی الگوها یا همبستگی‌های ظریف را آسان‌تر کنند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با تغییرات در الگوریتم‌های رمزنگاری یا اقدامات متقابل دفاعی سازگار شوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا ضعف‌های خاص در سیستم‌های رمزنگاری را با دقت بالا شناسایی کنند.
- **پنهان‌کاری:**
  - با استفاده از تحلیل‌های آماری و مدل‌های احتمالاتی، حملات مبتنی بر هوش مصنوعی می‌توانند حتی در سیستم‌های به‌خوبی ایمن‌شده نیز بدون تشخیص باقی بمانند.

1. Cryptanalysis

۲-۸-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های برچسب‌دار از متن‌های رمزنگاری‌شده و متن‌های آشکار متناظر آموزش می‌دهد تا ارتباط بین آن‌ها را یاد بگیرد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در متن‌های رمزنگاری‌شده بدون برچسب‌گذاری قبلی شناسایی می‌کند و به کشف ساختارهای پنهان کمک می‌کند.
  - یادگیری تقویتی: استراتژی‌های حمله را با پاداش دادن به تلاش‌های موفق رمزگشایی و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در متن‌های رمزنگاری‌شده را تحلیل می‌کنند و متن‌های آشکار یا کلیدهای احتمالی را پیش‌بینی می‌کنند.
  - شبکه‌های مولد تخاصمی متن‌های رمزنگاری‌شده مصنوعی اما واقع‌گرایانه برای اهداف آموزشی یا آزمایش استحکام طرح‌های رمزنگاری ایجاد می‌کنند.
- **پردازش زبان طبیعی:**
  - الگوهای زبانی در متن‌های آشکار رمزگشایی‌شده را تحلیل می‌کند تا نتایج را تأیید کرده و پیش‌بینی‌ها را بهبود بخشد.
- **تحلیل آماری:**
  - از نظریه احتمال و استنتاج آماری برای تخمین بیت‌های کلید یا بخش‌هایی از متن‌های آشکار بر اساس متن‌های رمزنگاری‌شده مشاهده‌شده استفاده می‌کند.
- **الگوریتم‌های تکاملی:**
  - از الگوریتم‌های ژنتیکی یا استراتژی‌های تکاملی برای بهبود تدریجی مدل‌های حمله با شبیه‌سازی فرآیندهای انتخاب طبیعی استفاده می‌کند.

۲-۸-۳ دارایی‌های هدف

- **الگوریتم‌های رمزنگاری متقارن:**
  - Blowfish، DES، AES و سایر رمزهای بلوکی که در برابر تحلیل الگو یا حملات بروت فورس تقویت‌شده توسط هوش مصنوعی آسیب‌پذیر هستند.

- **الگوریتم‌های رمزنگاری نامتقارن:**
  - RSA ،ECC و سایر سیستم‌های رمزنگاری کلید عمومی که در برابر حملات فاکتورگیری یا لگاریتم گسسته تسریع‌شده توسط یادگیری ماشین آسیب‌پذیر هستند.
- **توابع هش:**
  - SHA-1، MD5 و توابع هش قدیمی‌تر که در برابر حملات برخورد یا بازیابی پیش‌تصویر با استفاده از تکنیک‌های مبتنی بر هوش مصنوعی آسیب‌پذیر هستند.
- **رمزهای جریانی:**
  - ChaCha20، RC4 و رمزهای جریانی مشابه که در برابر پیش‌بینی جریان کلید یا بازیابی حالت آسیب‌پذیر هستند.
- **پروتکل‌های رمزنگاری:**
  - SSH، SSL/TLS و سایر پروتکل‌های ارتباطی که برای ایمن‌سازی داده‌ها در حال انتقال به رمزنگاری متکی هستند.
- **سیستم‌های بلاک‌چین:**
  - ارزهای دیجیتال و قراردادهای هوشمند که برای اعتبارسنجی تراکنش‌ها و مکانیزم‌های اجماع به یکپارچگی رمزنگاری وابسته هستند.

۴-۸-۲ آسیب‌پذیری‌ها

- **کلیدهای ضعیف:**
  - کلیدهای ضعیف یا قابل پیش‌بینی، موفقیت حملات مبتنی بر هوش مصنوعی را آسان‌تر می‌کنند.
- **اشکالات پیاده‌سازی:**
  - باگ‌ها یا تنظیمات نادرست در کتابخانه‌های رمزنگاری یا پیاده‌سازی‌های سخت‌افزاری، سیستم‌ها را در معرض سوءاستفاده قرار می‌دهند.
- **عدم وجود تصادفی‌سازی کافی:**
  - عدم وجود تصادفی‌سازی واقعی در تولید کلید یا مقادیر nonce، احتمال موفقیت رمزشکنی را افزایش می‌دهد.
- **نشت کانال جانبی:**
  - نشت زمان‌بندی، مصرف برق یا انتشارات الکترومغناطیسی، اطلاعات اضافی را در اختیار

مهاجمان قرار می‌دهد.

- **الگوریتم‌های منسوخ:**
  - استفاده از الگوریتم‌های منسوخ یا ناامن، سیستم‌ها را در برابر تکنیک‌های رمزشکنی پیشرفته آسیب‌پذیر می‌کند.

۲-۸-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم مجموعه‌داده‌های بزرگی از متن‌های رمزنگاری‌شده را از طریق رهگیری مستقیم یا تولید مصنوعی جمع‌آوری می‌کند.
- **پیش‌پردازش:**
  - مهاجم داده‌های جمع‌آوری‌شده را پاک‌سازی و نرمال‌سازی می‌کند و نویز را حذف کرده و ویژگی‌های مرتبط را بهبود می‌بخشد.
- **آموزش مدل:**
  - مهاجم مدل‌های یادگیری ماشین را بر روی مجموعه‌داده‌های برچسب‌دار آموزش داده تا ارتباط بین متن‌های رمزنگاری‌شده و متن‌های آشکار یا کلیدها را یاد بگیرند.
- **اجرای حمله:**
  - مهاجم مدل آموزش‌دیده را برای تحلیل متن‌های رمزنگاری‌شده در دنیای واقعی و استخراج اطلاعات حساس مستقر می‌کند.
- **پس‌پردازش:**
  - مهاجم داده‌های استخراج‌شده را با استفاده از تکنیک‌های پس‌پردازش بهبود داده تا دقت افزایش یابد و خطاها کاهش یابد.
- **سوءاستفاده:**
  - مهاجم از اسرار بازیابی‌شده برای اهداف مخرب، مانند رمزگشایی ارتباطات، جعل هویت کاربران یا سرقت دارایی‌های دیجیتال استفاده کنید.

۲-۸-۶ نشانه‌های احتمال نفوذ

- **الگوهای ترافیکی غیرعادی:**
  - ناهنجاری‌ها در ترافیک شبکه، مانند درخواست‌های مکرر برای داده‌های رمزنگاری‌شده یا حجم غیرعادی تبادل متن‌های رمزنگاری‌شده.

- **فعالیت محاسباتی افزایش‌یافته:**
  - استفاده بیشتر از CPU یا GPU مرتبط با تلاش‌های رمزشکنی.
- **تلاش‌های غیرمنتظره رمزگشایی:**
  - تلاش‌های ناموفق رمزگشایی به دنبال دسترسی موفق ممکن است نشان‌دهنده حملات بروت فورس یا مبتنی بر الگو باشد.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول کاربران، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **لاگ‌های سیستم:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مرتبط با عملیات رمزنگاری یا استفاده از منابع.

۲-۸-۷ تأثیرات

- **نقض محرمانگی:**
  - افشای داده‌های حساس، مانند اطلاعات شخصی، سوابق مالی یا مالکیت فکری.
- **ضررهای مالی:**
  - دسترسی غیرمجاز به سیستم‌های مالی یا سرقت دارایی‌های دیجیتال، مانند ارزهای دیجیتال.
- **اختلال در عملیات:**
  - اختلال در ارتباطات امن یا زیرساخت‌های حیاتی به دلیل به‌خطر افتادن سیستم‌های رمزنگاری.
- **آسیب به اعتبار:**
  - از دست دادن اعتماد مشتریان و ارزش برند پس از یک نقض.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، منجر به جریمه‌ها و پیامدهای قانونی می‌شود.
- **خطرات ایمنی:**
  - تهدید برای جان انسان‌ها در صورت به‌خطر افتادن دستگاه‌های پزشکی، سیستم‌های حمل‌ونقل یا کنترل‌های صنعتی.

۲-۸-۸ راه‌های مقابله

- **استفاده از استانداردهای رمزنگاری قوی:**
  - از الگوریتم‌های رمزنگاری مدرن و به‌خوبی آزمایش‌شده مانند RSA-4096، AES-256 یا ECC با اندازه کلید کافی استفاده کنید.
- **مدیریت صحیح کلیدها:**
  - روش‌های ایمن تولید، ذخیره‌سازی و چرخش کلیدها را پیاده‌سازی کنید تا خطر به‌خطر افتادن کلیدها به حداقل برسد.
- **تولید اعداد تصادفی:**
  - اطمینان حاصل کنید که از مولدهای اعداد تصادفی رمزنگاری‌شده (CSPRNGs) برای تمام مقادیر تصادفی استفاده می‌شود.
- **به‌روزرسانی‌های منظم:**
  - کتابخانه‌ها و سیستم‌های رمزنگاری را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده وصله شوند.
- **محافظت در برابر کانال‌های جانبی:**
  - اقدامات متقابل مانند ماسک‌کردن، کورسازی یا تزریق نویز را برای مقابله با حملات کانال جانبی پیاده‌سازی کنید.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت که نشان‌دهنده تلاش‌های رمزشکنی هستند، استفاده کنید.
- **امنیت چندلایه:**
  - رمزنگاری را با سایر اقدامات امنیتی، مانند فایروال‌ها، سیستم‌های تشخیص نفوذ (IDS) و کنترل‌های دسترسی ترکیب کنید.
- **بررسی‌های منظم:**
  - ممیزی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌های بالقوه در پیاده‌سازی‌های رمزنگاری را شناسایی و برطرف کنید.

۲-۹ کش‌کاوی

حمله کش‌کاوی[1] شامل سوءاستفاده از رفتار کش یک سیستم برای استخراج اطلاعات حساس، مانند

1. Cache Mining Attack

کلیدهای رمزنگاری، رمزهای عبور یا داده‌های شخصی است. در حمله کش‌کاوی مبتنی بر هوش مصنوعی، از هوش مصنوعی برای افزایش دقت، کارایی و پنهان‌کاری حملات سنتی مبتنی بر کش مانند حملات زمان‌بندی، حملات کانال جانبی یا حملات برخورد کش استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند الگوهای استفاده از کش را تحلیل کنند، رفتارهای دسترسی به حافظه را پیش‌بینی کنند و داده‌های حساس را با دقت بیشتری استنباط کنند.

حملات کش‌کاوی به‌ویژه در محیط‌های ابری، منابع محاسباتی اشتراکی یا سیستم‌هایی که چندین فرآیند از سلسله مراتب کش یکسانی استفاده می‌کنند، خطرناک هستند. این حملات می‌توانند با تمرکز بر متاداده‌ها یا سیگنال‌های غیرمستقیم به جای دسترسی مستقیم به داده‌های محافظت‌شده، از رمزنگاری و سایر اقدامات امنیتی عبور کنند.

۲-۹-۱ ویژگی‌های کلیدی

- **تحلیل مبتنی بر داده:**
  - هوش مصنوعی حجم عظیمی از داده‌های مرتبط با کش، مانند زمان‌های دسترسی، الگوهای حذف و آدرس‌های حافظه را پردازش می‌کند تا همبستگی‌های بین ورودی‌ها و خروجی‌ها را کشف کند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند شناسایی آسیب‌پذیری‌ها را خودکار می‌کنند و تلاش دستی را کاهش داده و سرعت سوءاستفاده را افزایش می‌دهند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند به‌طور پویا با تغییرات در سیستم هدف، مانند الگوریتم‌های به‌روزرسانی‌شده، تغییرات سخت‌افزاری یا مکانیزم‌های دفاعی سازگار شوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا بیت‌ها یا بایت‌های خاصی از داده‌های حساس را با دقت بالا شناسایی کنند.
- **پنهان‌کاری:**
  - با استفاده از سیگنال‌های ظریفی که اغلب نادیده گرفته می‌شوند، حملات کش‌کاوی مبتنی بر هوش مصنوعی می‌توانند حتی در محیط‌های به‌خوبی ایمن‌سازی‌شده نیز کشف نشوند.

۲-۹-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را آموزش می‌دهد تا الگوهای زمان‌های دسترسی به کش، الگوهای حذف یا استفاده از حافظه را تشخیص دهند و آن‌ها را به مقادیر محرمانه مربوطه (مانند کلیدهای رمزنگاری) نگاشت کنند.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در رفتار کش را بدون برچسب‌گذاری قبلی شناسایی می‌کند.
  - یادگیری تقویتی: استراتژی حمله را با پاداش دادن به استخراج موفقیت‌آمیز داده و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در داده‌های سری زمانی، مانند زمان‌های دسترسی به کش یا توالی‌های حذف را تحلیل می‌کنند.
  - شبکه‌های مولد تخاصمی: الگوهای دسترسی به کش مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا برای اهداف آموزشی یا آزمایش استحکام اقدامات دفاعی استفاده شوند.
- **تحلیل سری‌های زمانی:**
  - از روش‌های آماری و یادگیری ماشین برای تحلیل وابستگی‌های زمانی در الگوهای دسترسی به کش استفاده می‌کند تا رفتارهای آینده را پیش‌بینی کند.
- **پردازش سیگنال:**
  - تکنیک‌هایی مانند تبدیل‌های فوریه، تحلیل موجک یا تحلیل مؤلفه‌های اصلی (PCA) را برای پیش‌پردازش و استخراج ویژگی‌های معنادار از داده‌های خام مرتبط با کش اعمال می‌کند.

۲-۹-۳ دارایی‌های هدف

- **خدمات ابری:**
  - ماشین‌های مجازی (VM)[1] و کانتینرهایی که منابع را در پلتفرم‌های چندمستأجری به اشتراک می‌گذارند، جایی که حملات زمان‌بندی یا مبتنی بر کش می‌توانند اطلاعات حساس را فاش کنند.

1. Virtual Machine (VM)

- **سیستم‌های رمزنگاری:**
  - ماژول‌های امنیتی سخت‌افزاری (HSM)[1]، کارت‌های هوشمند و دستگاه‌های تعبیه‌شده که الگوریتم‌های رمزنگاری را پیاده‌سازی می‌کنند و به الگوهای دسترسی به حافظه قابل پیش‌بینی متکی هستند.
- **سیستم‌های سازمانی:**
  - شبکه‌های شرکتی، سرورهای فایل و سیستم‌های پایگاه داده که داده‌های حیاتی کسب‌وکار را ذخیره می‌کنند، جایی که کش‌های اشتراکی ممکن است اطلاعات حساس را فاش کنند.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های IoT با منابع محدود که محافظت‌های کمی در برابر حملات کش‌کاوی دارند.
- **سیستم‌های پرداخت:**
  - ترمینال‌های فروش، دستگاه‌های خواننده کارت و سیستم‌های پرداخت بدون تماس که در برابر نشت اطلاعات تراکنش‌ها از طریق کش آسیب‌پذیر هستند.

۴-۹-۲ آسیب‌پذیری‌ها

- **منابع اشتراکی:**
  - محیط‌های چندمستأجری، مانند پلتفرم‌های ابری، به مهاجمان امکان می‌دهند تا از کش‌ها، حافظه یا پردازنده‌های اشتراکی برای استنباط اطلاعات حساس سوءاستفاده کنند.
- **الگوهای دسترسی قابل پیش‌بینی:**
  - عملیات رمزنگاری یا سایر فرآیندهای بحرانی امنیتی سیگنال‌های قابل اندازه‌گیری، مانند تغییرات زمان‌بندی یا حذف کش، منتشر می‌کنند که می‌توانند تحلیل شوند.
- **اقدامات تدافعی ضعیف:**
  - تکنیک‌های ناکافی پوشش، کورسازی یا تزریق نویز، سیستم‌ها را در برابر حملات کش‌کاوی آسیب‌پذیر می‌کند.
- **خطای انسانی:**
  - انتخاب‌های طراحی ضعیف، پیکربندی‌های نادرست یا تست ناکافی در طول توسعه ممکن است آسیب‌پذیری‌هایی را ایجاد کند که از طریق کش‌کاوی قابل سوءاستفاده باشند.

1. Hardware Security Module (HSM)

- **عدم نظارت:**
  - عدم نظارت مداوم بر فعالیت‌های غیرعادی کش، به حملات امکان می‌دهد تا بدون تشخیص پیش‌بروند.

۲-۹-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای تخصصی یا ابزارهای ابزاردقیق برای نظارت بر زمان‌های دسترسی به کش، الگوهای حذف یا استفاده از حافظه استفاده می‌کند.
- **پیش‌پردازش:**
  - مهاجم تکنیک‌های پردازش سیگنال را برای پاک‌سازی و نرمال‌سازی داده‌های جمع‌آوری‌شده اعمال می‌کند، نویز را حذف کرده و ویژگی‌های مرتبط را بهبود می‌بخشد.
- **آموزش مدل:**
  - مهاجم مدل‌های یادگیری ماشین را بر روی مجموعه‌های داده برچسب‌خورده آموزش می‌دهد تا رفتار کش را با مقادیر محرمانه (مانند کلیدهای رمزنگاری) مرتبط کند.
- **اجرای حمله:**
  - مهاجم مدل آموزش‌دیده را برای تحلیل فعالیت کش در زمان واقعی و استخراج اطلاعات حساس مستقر می‌کند.
- **پس‌پردازش:**
  - مهاجم داده‌های استخراج‌شده را با استفاده از تکنیک‌های پس‌پردازش بهبود می‌بخشد تا دقت را افزایش داده و خطاها را کاهش دهد.
- **سوءاستفاده:**
  - مهاجم از اسرار بازیابی‌شده برای اهداف مخرب، مانند رمزگشایی ارتباطات حساس یا جعل کاربران قانونی استفاده می‌کند.

۲-۹-۶ نشانه‌های احتمال نفوذ

- **فعالیت غیرعادی کش:**
  - ناهنجاری‌ها در الگوهای دسترسی به کش، مانند توالی‌های حذف غیرمنتظره یا افزایش رقابت.

- **تغییرات زمان‌بندی:**
  - انحرافات در زمان‌های اجرای فرآیندهای بحرانی امنیتی که با عملیات یا رویدادهای خاص همبستگی دارند.
- **ورودی‌های غیرمنتظره در لاگ‌ها:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به استفاده از حافظه، مدیریت کش یا فعالیت‌های غیرعادی.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول سیستم ناسازگار است، مانند تلاش‌های مکرر برای دسترسی به منابع یا عملکردهای خاص.

۲-۹-۷ تأثیرات

- **به‌خطر افتادن محرمانگی:**
  - افشای داده‌های حساس، مانند کلیدهای رمزنگاری، رمزهای عبور یا اطلاعات شخصی.
- **خسارات مالی:**
  - دسترسی غیرمجاز به سیستم‌های مالی یا سرقت دارایی‌های دیجیتال.
- **اختلال در عملیات:**
  - اختلال در ارتباطات امن یا زیرساخت‌های حیاتی به دلیل به‌خطر افتادن سیستم‌های رمزنگاری.
- **آسیب به اعتبار:**
  - از دست دادن اعتماد مشتری و ارزش برند پس از یک نفوذ.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

۲-۹-۸ راه‌های مقابله

- **تقسیم‌بندی کش:**
  - فرآیندهای حساس یا مستأجران را در بخش‌های جداگانه کش ایزوله کنید تا از تداخل جلوگیری شود.

- **تصادفی‌سازی:**
  - تصادفی‌سازی را در الگوهای دسترسی به حافظه معرفی کنید تا عملیات حساس را مبهم کرده و تحلیل آن‌ها را دشوارتر کنید.
- **تزریق نویز:**
  - نویز کنترل‌شده را به فعالیت کش اضافه کنید تا توانایی مهاجمان برای استنباط اطلاعات مفید را مختل کنید.
- **روش‌های کدنویسی ایمن:**
  - الگوریتم‌های زمان‌ثابت را پیاده‌سازی کنید و از شاخه‌های شرطی که می‌توانند اطلاعات زمان‌بندی را فاش کنند، اجتناب کنید.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌های بالقوه کش‌کاوی را شناسایی و برطرف کنید.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت کش و ایجاد هشدار استفاده کنید.
- **ایزوله‌سازی منابع:**
  - اطمینان حاصل کنید که منابع در محیط‌های چندمستأجری به‌درستی ایزوله شده‌اند تا از حملات cross-VM یا cross-container جلوگیری شود.
- **اقدامات امنیتی فیزیکی:**
  - سیستم‌های حیاتی را با موانع فیزیکی، محافظها یا محفظه‌های مقاوم در برابر دستکاری محافظت کنید تا دسترسی به سیگنال‌های مرتبط با کش محدود شود.

۲-۱۰ مرد میانی

حمله مرد میانی (MitM)[1] شامل یک مهاجم است که ارتباطات بین دو طرف را بدون اطلاع آنها رهگیری و احتمالاً تغییر می‌دهد. در حمله مرد میانی مبتنی بر هوش مصنوعی، از هوش مصنوعی برای افزایش دقت، انعطاف‌پذیری و پنهان‌کاری تکنیک‌های سنتی MitM استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند ترافیک رمزنگاری‌شده را تحلیل کنند، الگوهای ارتباطی را

1. Man-in-the-Middle Attack

پیش‌بینی کرده و فرآیند رهگیری و دستکاری را به صورت خودکار و در مقیاس بزرگ انجام دهند.

این نوع حمله به یکپارچگی و محرمانه‌بودن ارتباطات حمله می‌کند و به مهاجمان اجازه می‌دهد اطلاعات حساس را سرقت کنند، بارهای مخرب[1] تزریق کنند یا تراکنش‌های حیاتی را مختل نمایند. حملات MitM مبتنی بر هوش مصنوعی می‌توانند با بهره‌برداری از نقاط ضعف پروتکل‌ها، رفتار انسانی یا پیکربندی سیستم‌ها، دفاع‌های سنتی مانند رمزنگاری، فایروال‌ها یا سیستم‌های تشخیص ناهنجاری را دور بزنند.

۲-۱۰-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی شناسایی کانال‌های ارتباطی آسیب‌پذیر را خودکار می‌کند، تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی به‌طور پویا تکامل می‌یابند و با تغییرات در الگوریتم‌های رمزنگاری، پروتکل‌های شبکه یا اقدامات دفاعی سازگار می‌شوند.
- **پنهان‌کاری:**
  - با تقلید از الگوهای ترافیک قانونی یا رمزنگاری داده‌های رهگیری‌شده تا زمان تحلیل، هوش مصنوعی اطمینان می‌دهد که فعالیت‌های MitM کشف نشوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان اجازه می‌دهند کانال‌های ارتباطی خاص یا اهداف باارزش را با دقت بیشتری شناسایی کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد ارتباطات چندین سیستم یا شبکه را به‌طور همزمان رهگیری و دستکاری کنند.

۲-۱۰-۱ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های حملات MitM موفق و ناموفق آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.

1. payload

- یادگیری بدون نظارت: الگوها یا ناهنجاری‌ها در ترافیک شبکه را شناسایی می‌کند تا آسیب‌پذیری‌های بالقوه را آشکار کند.
- یادگیری تقویتی: تاکتیک‌های رهگیری و دستکاری را با پاداش‌دهی به نتایج موفق و جریمه‌کردن شکست‌ها بهینه می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در ترافیک رمزنگاری‌شده یا رفتار پروتکل‌ها را تحلیل می‌کنند تا نقاط ضعف را شناسایی کنند.
  - شبکه‌های مولد تخاصمی: الگوهای ترافیک مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های فرار را آزمایش و بهبود بخشند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی در ارتباطات رهگیری‌شده را تحلیل می‌کند تا زمینه‌های بیشتری درباره آسیب‌پذیری‌های بالقوه یا مکانیسم‌های دفاعی استخراج کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا فعالیت‌های مخرب را با فعالیت‌های قانونی ترکیب کند.
- **تحلیل ترافیک:**
  - از هوش مصنوعی برای تحلیل فراداده‌ها، زمان‌بندی یا اندازه بسته‌ها استفاده می‌کند تا اطلاعات حساس را حتی در صورت رمزنگاری محتوا استنباط کند.

۳-۱۰-۱ دارایی‌های هدف

- **برنامه‌های وب:**
  - پلتفرم‌های تجارت الکترونیک، خدمات مالی یا پورتال‌های بهداشتی که در آن‌ها رهگیری ارتباطات می‌تواند منجر به نقض داده یا کلاهبرداری شود.
- **شبکه‌های سازمانی:**
  - اینترانت‌های شرکتی یا محیط‌های کار از راه دور که در آن‌ها حملات MitM می‌تواند ارتباطات حساس کسب‌وکار را به خطر بیندازد.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های IoT با منابع محدود که اغلب از رمزنگاری سبک یا پروتکل‌های ارتباطی ناامن

استفاده می‌کنند.

- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، سیستم‌های حمل‌ونقل یا سیستم‌های کنترل صنعتی که در آن‌ها حملات MitM می‌تواند باعث اختلال در عملیات شود.
- **ارتباطات موبایل:**
  - رهگیری ترافیک برنامه‌های موبایل، پیام‌های متنی یا تماس‌های VoIP برای سرقت اعتبارنامه‌ها، استراق سمع یا تزریق محتوای مخرب.

۲-۱۰-۴ آسیب‌پذیری‌ها

- **رمزنگاری ضعیف:**
  - استفاده از پروتکل‌های رمزنگاری قدیمی یا نادرست، ارتباطات را در معرض رهگیری و رمزگشایی قرار می‌دهد.
- **احراز هویت ناکافی:**
  - عدم وجود مکانیسم‌های احراز هویت متقابل قوی، به مهاجمان اجازه می‌دهد تا در طول ارتباط، هویت نهادهای مورد اعتماد را جعل کنند.
- **خطای انسانی:**
  - پیکربندی نادرست فایروال‌ها، گواهی‌های مشترک یا آموزش ناکافی به‌طور ناخواسته آسیب‌پذیری‌هایی را ایجاد می‌کند که توسط حملات MitM قابل بهره‌برداری هستند.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های قدیمی یا نرم‌افزارهای منسوخ، سازمان‌ها را در برابر تکنیک‌های پیشرفته MitM آسیب‌پذیر می‌کند.
- **شبکه‌های وای فای عمومی:**
  - شبکه‌های عمومی ناامن یا کم‌امن، زمینه‌های مناسبی برای حملات MitM مبتنی بر هوش مصنوعی فراهم می‌کنند.

۲-۱۰-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای نقشه‌برداری از شبکه هدف، شناسایی کانال‌های ارتباطی و ارزیابی قدرت رمزنگاری استفاده می‌کند.

- **اسکن آسیب‌پذیری:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل شبکه و شناسایی نقاط ضعف، مانند اعتبارسنجی نادرست گواهی‌ها یا رمزنگاری ضعیف، به‌کار می‌گیرد.
- **تنظیم رهگیری:**
  - مهاجم، گره مهاجم را به‌طور استراتژیک در شبکه قرار می‌دهد تا ارتباطات را از طریق جعل ARP[1] یا ربودن سیستم نام دامنه (DNS) رهگیری کند.
- **تلاش برای رمزگشایی:**
  - مهاجم از هوش مصنوعی برای تحلیل ترافیک رمزنگاری‌شده، پیش‌بینی کلیدها یا بهره‌برداری از نقاط ضعف پروتکل‌ها برای رمزگشایی داده‌های رهگیری‌شده استفاده می‌کند.
- **اجرای دستکاری:**
  - مهاجم ارتباطات رهگیری‌شده را در زمان واقعی تغییر می‌دهد و بارهای مخرب تزریق می‌کند یا جزئیات تراکنش‌ها را تغییر می‌دهد.
- **پایداری:**
  - مهاجم حمله را بر اساس بازخورد به‌طور مداوم بهبود می‌بخشد تا دسترسی بلندمدت و تأثیر بر ارتباطات را تضمین کند.

۶-۱۰-۲ نشانه‌های احتمال نفوذ

- **الگوهای ترافیک غیرعادی:**
  - ناهنجاری‌ها در اندازه بسته‌ها، زمان‌های بین ورود یا جهت جریان‌ها که از رفتار پایه منحرف می‌شوند.
- **افزایش تاخیر:**
  - افزایش ناگهانی تاخیر در ارتباطات که ممکن است نشان‌دهنده رهگیری یا دستکاری باشد.
- **هشدارهای گواهی:**
  - خطاهای غیرمنتظره یا عدم تطابق گواهی‌های SSL/TLS که نشان‌دهنده فعالیت احتمالی MitM است.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار عادی کاربر یا سیستم، مانند دسترسی به مناطق محدود یا انجام

1. Address Resolution Protocol (ARP)

پرس‌وجوهای غیرعادی.

- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.

۲-۱۰-۷ تأثیرات

- **نقض داده:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به ضرر مالی یا آسیب به اعتبار می‌شود.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، فعالیت‌های کلاهبرداری یا اختلال در خدمات که باعث آسیب مالی می‌شوند.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل ارتباطات مخدوش یا تزریق بدافزار.
- **آسیب به اعتبار:**
  - از دست دادن اعتماد مشتریان و ارزش برند پس از نقض یا سوءاستفاده از داده‌های سرقت‌شده.
- **فساد سیستم:**
  - تزریق بارهای مخرب به ارتباطات رهگیری‌شده می‌تواند پایگاه‌های داده، برنامه‌ها یا فریم‌ور را فاسد کند.

۲-۱۰-۸ راه‌های مقابله

- **رمزنگاری انتها به انتها:**
  - پروتکل‌های رمزنگاری قوی مانند TLS 1.3 یا رمزنگاری انتها به انتها را پیاده‌سازی کنید تا ارتباطات از رهگیری محافظت شوند.
- **احراز هویت متقابل:**
  - از هر دو طرف در یک ارتباط بخواهید هویت یکدیگر را تأیید کنند تا از جعل هویت نهادهای مورد اعتماد جلوگیری شود.

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های IDS/IPS مبتنی بر هوش مصنوعی را مستقر کنید که قادر به تشخیص ناهنجاری‌های ظریف در الگوهای ترافیک هستند که نشان‌دهنده حملات MitM است.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای نظارت بر انحرافات در رفتار ارتباطی استفاده کنید که ممکن است نشان‌دهنده قصد مخرب باشد.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فریم‌ورها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده در حملات MitM اصلاح شوند.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش دسترسی غیرمجاز یا دستکاری داده‌ها محدود شود.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد شناسایی علائم حملات MitM، مانند هشدارهای گواهی یا تاخیر غیرعادی، آموزش دهید.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور MitM و تهدیدات انطباق‌پذیر مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حملات MitM را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.
- **پیاده‌سازی DNSSEC:**
  - از DNSSEC (امنیت گسترش‌یافته سیستم نام دامنه) استفاده کنید تا با احراز هویت پاسخ‌های سیستم نام دامنه (DNS)، از حملات MitM مبتنی بر DNS جلوگیری کنید.

۲-۱۱ تولید خودکار دامنه

الگوریتم‌های تولید دامنه (DGAs)[1] توسط بدافزارها برای تولید پویای تعداد زیادی از نام‌های دامنه به

1. Domain Generation Algorithm (DGA)

منظور ارتباطات فرمان و کنترل (C2) استفاده می‌شوند. در یک حمله تولید خودکار دامنه‌های مبتنی بر هوش مصنوعی، هوش مصنوعی با استفاده از یادگیری ماشین و یادگیری عمیق، تکنیک‌های سنتی DGA را بهبود می‌بخشد تا نام‌های دامنه‌ای بسیار غیرقابل پیش‌بینی، آگاه از زمینه و تطبیق‌پذیر ایجاد کند. این کار باعث می‌شود که دفاع در برابر این دامنه‌های مخرب به طور قابل توجهی سخت‌تر شود و مهاجمان بتوانند ارتباطات پایدار با سیستم‌های آلوده حفظ کنند.

تولید دامنه‌های مبتنی بر هوش مصنوعی به ویژه خطرناک هستند زیرا می‌توانند مکانیسم‌های سنتی مانند لیست‌های سیاه، سینک‌هولینگ و تشخیص ناهنجاری را دور بزنند و به مهاجمان اجازه دهند از تشخیص فرار کرده و کمپین‌های بلندمدت خود را حفظ کنند.

۲-۱۱-۱ ویژگی‌های کلیدی

- **غیرقابل پیش‌بینی بودن بالا:**
  - هوش مصنوعی نام‌های دامنه‌ای تولید می‌کند که تقریباً غیرممکن است با استفاده از روش‌های سنتی مبتنی بر الگو پیش‌بینی شوند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند تولید و حل نام‌های دامنه را خودکار می‌کنند، که باعث کاهش تلاش دستی و افزایش کارایی می‌شود.
- **تطبیق‌پذیری:**
  - تولید دامنه‌های مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و خود را با تغییرات در اقدامات دفاعی مانند فیلتر سیستم نام دامنه (DNS) یا مسدودسازی مبتنی بر اعتبار تطبیق دهند.
- **پنهان‌کاری:**
  - دامنه‌های تولید شده از قواعد نام‌گذاری قانونی یا روندها تقلید می‌کنند، که تشخیص آن‌ها از ترافیک بی‌خطر را سخت‌تر می‌کند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی امکان تولید سریع میلیون‌ها نام دامنه را فراهم می‌کند و مکانیسم‌های دفاعی سنتی را تحت فشار قرار می‌دهد.

۲-۱۱-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها بر روی مجموعه‌داده‌هایی از دامنه‌های قانونی و مخرب شناخته

شده آموزش می‌بینند تا الگوها را یاد بگیرند و دامنه‌های جدیدی تولید کنند که با محیط ترکیب شوند.

- یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در ساختار نام‌های دامنه را شناسایی می‌کند تا استراتژی‌های تولید را بدون برچسب‌گذاری قبلی بهبود بخشد.
- یادگیری تقویتی: تولید دامنه را با پاداش دادن به موفقیت‌ها در حل دامنه و جریمه کردن تلاش‌های مسدود یا علامت‌گذاری شده بهینه می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، الگوهای زبانی پیچیده در نام‌های دامنه را تحلیل می‌کنند تا نتایج واقعی و آگاه از زمینه تولید کنند.
  - شبکه‌های تولیدی تخاصمی: دامنه‌های مصنوعی اما متقاعدکننده ایجاد می‌کنند که قابلیت‌های فرار از تشخیص را آزمایش و بهبود می‌بخشند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی مانند موضوعات داغ، نام‌های برند یا اصطلاحات صنعتی را تحلیل می‌کند تا دامنه‌هایی ایجاد کند که با علایق یا رویدادهای فعلی هماهنگ باشند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا مهاجمان بتوانند دامنه‌هایی تولید کنند که با الگوهای مورد انتظار مطابقت داشته باشند.
- **تحلیل زمانی:**
  - از تحلیل سری‌های زمانی برای پیش‌بینی روندهای آینده در قواعد نام‌گذاری دامنه استفاده می‌کند و خود را با آن‌ها تطبیق می‌دهد.

۳-۱۱-۲ دارایی‌های هدف

- **ارتباطات فرمان و کنترل (C2) بدافزار:**
  - مکانیسمی مقاوم برای دستگاه‌های آلوده فراهم می‌کند تا با سرورهای کنترل شده توسط مهاجم ارتباط برقرار کنند و کنترل پایدار را تضمین کنند.
- **شبکه‌های سازمانی:**
  - شبکه‌های شرکتی که در آن‌ها مسدود کردن دامنه‌های مخرب برای جلوگیری از خروج داده‌ها یا آلودگی بیشتر حیاتی است.

- **زیرساخت سیستم نام دامنه (DNS):**
  - از ضعف‌های فرآیندهای حل سیستم نام دامنه (DNS) برای حل دامنه‌های تولید شده و ایجاد کانال‌های فرمان و کنترل (C2) سوءاستفاده می‌کند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تاسیسات بهداشتی، موسسات مالی یا سیستم‌های حمل و نقل که در آن‌ها حفظ ارتباطات پنهانی می‌تواند منجر به آسیب‌های قابل توجه شود.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های IoT با منابع محدود که اغلب به عنوان بخشی از بات‌نت‌ها استفاده می‌شوند، و در آن‌ها DGA ارتباط مطمئن با سرورهای C2 را تضمین می‌کند.

۴-۱۱-۲ آسیب‌پذیری‌ها

- **فیلتر DNS ضعیف:**
  - عدم وجود مکانیسم‌های قوی فیلتر سیستم نام دامنه (DNS) باعث می‌شود دامنه‌های تولید شده توسط هوش مصنوعی بدون تشخیص عبور کنند.
- **لیست‌های سیاه ناکافی:**
  - لیست‌های سیاه سنتی نمی‌توانند با حجم و غیرقابل پیش‌بینی بودن DGAهای مبتنی بر هوش مصنوعی همگام شوند.
- **خطای انسانی:**
  - فایروال‌ها، سیاست‌های سیستم نام دامنه (DNS) یا ابزارهای امنیتی نادرست پیکربندی شده، به طور ناخواسته آسیب‌پذیری‌هایی را در معرض سوءاستفاده توسط الگوریتم تولید داده (DGAs) قرار می‌دهند.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های تشخیص قدیمی یا نادرست باعث می‌شود سازمان‌ها در برابر DGAهای تطبیق‌پذیر آسیب‌پذیر باشند.
- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - نقص‌های کشف نشده در زیرساخت DNS یا پروتکل‌ها، نقاط ورودی را برای DGAهای مبتنی بر هوش مصنوعی فراهم می‌کنند تا کانال‌های ارتباطی ایجاد کنند.

۲-۱۱-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره قواعد نام‌گذاری قانونی دامنه، روندها و رفتارها استفاده می‌کند.
- **آموزش مدل:**
  - مدل‌های یادگیری ماشین را بر روی مجموعه‌داده‌های برچسب‌گذاری شده از دامنه‌های قانونی و مخرب آموزش می‌دهد تا الگوها را یاد بگیرند و دامنه‌های جدیدی تولید کنند.
- **تولید دامنه:**
  - مدل آموزش دیده را برای تولید مجموعه‌ای بزرگ از نام‌های دامنه که از طرح‌های نام‌گذاری قانونی تقلید می‌کنند و از تشخیص فرار می‌کنند، به کار می‌گیرد.
- **آزمایش و اصلاح:**
  - دامنه‌های تولید شده را در برابر دفاع‌های شبیه‌سازی شده آزمایش می‌کند تا اطمینان حاصل کند که از تشخیص فرار می‌کنند و الگوریتم را بر این اساس اصلاح می‌کند.
- **استقرار:**
  - از دامنه‌های تولید شده برای ارتباطات فرمان و کنترل (C2) استفاده می‌کند تا اطمینان حاصل کند دستگاه‌های آلوده می‌توانند آن‌ها را حل کرده و با مهاجم در ارتباط بمانند.
- **پایداری:**
  - به طور مداوم لیست دامنه‌های تولید شده را به‌روزرسانی می‌کند تا با تغییرات در اقدامات دفاعی تطبیق یابد و ارتباطات را حفظ کند.

۲-۱۱-۶ نشانه‌های احتمال نفوذ

- **درخواست‌های DNS غیرعادی:**
  - ناهنجاری‌ها در الگوهای پرس‌وجوی DNS، مانند حجم بالای درخواست‌ها برای دامنه‌های غیرموجود یا به ندرت دسترسی‌پذیر.
- **قواعد نام‌گذاری تصادفی:**
  - دامنه‌هایی با نام‌های به ظاهر تصادفی یا بی‌معنی که از قواعد نام‌گذاری معمول منحرف می‌شوند.

- **افزایش تلاش‌های حل:**
  - تلاش‌های مکرر ناموفق برای حل دامنه‌های خاص، که نشان‌دهنده استفاده از DGA است.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار عادی شبکه ناسازگار هستند، مانند اتصالات خروجی غیرمنتظره یا فعالیت‌های DNS غیرعادی.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های DNS مربوط به پرس‌وجوهای دامنه‌های حل نشده یا علامت‌گذاری شده.

۲-۱۱-۷ تأثیرات

- **تهدیدات پایدار:**
  - دسترسی بلندمدت به سیستم‌های آلوده را حفظ می‌کند و امکان خروج داده‌ها یا کنترل از راه دور را فراهم می‌کند.
- **اختلال عملیاتی:**
  - خدمات یا سیستم‌های حیاتی کسب‌وکار را به دلیل سوءاستفاده از اعتبارنامه‌ها یا آسیب‌پذیری‌ها مختل می‌کند.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلالات خدمات که باعث آسیب‌های مالی می‌شوند.
- **جریمه‌های نظارتی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.
- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا فریم‌ورهای حیاتی، که سیستم‌ها را غیرقابل استفاده می‌کند.

۲-۱۱-۸ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به شناسایی الگوهای

پرس‌وجوی DNS ناهنجار مرتبط با DGAهای مبتنی بر هوش مصنوعی هستند.

- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص فعالیت‌های DNS غیرعادی یا الگوهای ارتباطی مرتبط با الگوریتم تولید داده (DGAs) استفاده کنید.
- **فیلتر DNS:**
  - راه‌حل‌های پیشرفته فیلتر DNS را پیاده‌سازی کنید که شامل اطلاعات تهدیدات بلادرنگ و یادگیری ماشین برای مسدود کردن دامنه‌های مخرب هستند.
- **لیست‌های سیاه پیش‌گیرانه:**
  - از تحلیل‌های پیش‌بینی‌کننده برای مسدود کردن پیش‌دستانه دامنه‌های تولید شده توسط DGA بر اساس الگوهای مشاهده شده استفاده کنید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش ترافیک مخرب را محدود کرده و تاثیر حملات موفق را کاهش دهید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فریم‌ورها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته شده که توسط DGAs سوءاستفاده می‌شوند، رفع شوند.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد شناسایی نشانه‌های فعالیت‌های مخرب، مانند پرس‌وجوهای DNS غیرعادی یا اتصالات غیرمنتظره، آموزش دهید.
- **اطلاعات تهدیدات پیشرفته:**
  - از پلتفرم‌های اطلاعات تهدیدات مبتنی بر هوش مصنوعی استفاده کنید تا درباره تکنیک‌های نوظهور DGA و تهدیدات تطبیق‌پذیر آگاه بمانید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به سرعت حملات مبتنی بر DGA را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.
- **پیاده‌سازی DNSSEC:**
  - DNSSEC (امنیت گسترش‌یافته سیستم نام دامنه) را مستقر کنید تا پاسخ‌های DNS را احراز

هویت کرده و از جعل یا تغییر مسیر جلوگیری کنید.

۲-۱۲ مبهم‌سازی بدافزار

مبهم‌سازی بدافزار[1] شامل پنهان‌کردن نرم‌افزارهای مخرب برای فرار از تشخیص توسط برنامه‌های ضدویروس، سیستم‌های تشخیص نفوذ (IDS) و سایر اقدامات امنیتی است. در حمله مبهم‌سازی بدافزار مبتنی بر هوش مصنوعی، از هوش مصنوعی برای خودکارسازی و بهبود فرآیند پنهان‌کردن امضاها، رفتارها یا ساختارهای کد بدافزار استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند بدافزارهای بسیار چندشکلی، متامورفیک یا پنهان‌کاری ایجاد کنند که به‌طور پویا برای اجتناب از تشخیص سازگار می‌شوند.

این نوع حمله به‌ویژه خطرناک است زیرا به بدافزار اجازه می‌دهد تا حتی از راه‌حل‌های امنیتی پیشرفته مانند بخش‌های ایزوله، تحلیل اکتشافی یا نظارت رفتاری عبور کند. مبهم‌سازی بدافزار مبتنی بر هوش مصنوعی به مهاجمان امکان می‌دهد تا تهدیدات خود را به‌طور مداوم تکامل دهند و تشخیص و تحلیل آن‌ها را دشوارتر کنند.

۲-۱۲-۱ ویژگی‌های کلیدی

- **تبدیل پویای کد:**
  - هوش مصنوعی ساختار، منطق یا ظاهر بدافزار را در زمان واقعی تغییر می‌دهد تا از تشخیص مبتنی بر امضا اجتناب کند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند ایجاد انواع جدید بدافزار را خودکار می‌کنند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهند.
- **انعطاف‌پذیری:**
  - بدافزار مبتنی بر هوش مصنوعی می‌تواند با گذشت زمان تکامل یابد و با تغییرات در اقدامات امنیتی، فناوری‌های دفاعی یا شرایط محیطی سازگار شود.
- **پنهان‌کاری:**
  - با تقلید از رفتار نرم‌افزارهای قانونی یا رمزنگاری payloadها تا زمان اجرا، بدافزار مبتنی بر هوش مصنوعی برای مدت‌های طولانی کشف نشده باقی می‌ماند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا تعداد زیادی نمونه بدافزار منحصر به فرد

1. malware obfuscation

تولید کنند و مکانیزم‌های تشخیص سنتی را تحت فشار قرار دهند.

۲-۱۲-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌های داده‌ای از بدافزارهای شناخته‌شده آموزش می‌دهد تا الگوها را یاد بگیرد و انواع جدیدی با امضاهای تغییر یافته ایجاد کند.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در کد بدافزار شناسایی می‌کند تا تکنیک‌های فرار بالقوه را کشف کند.
  - یادگیری تقویتی: استراتژی‌های مبهم‌سازی را با پاداش دادن به فرار موفقیت‌آمیز و جریمه کردن تشخیص بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در کد باینری، ترافیک شبکه یا رفتار سیستم را تحلیل می‌کنند تا تکنیک‌های مبهم‌سازی را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: نمونه‌های بدافزار مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا برای اهداف آموزشی یا آزمایش اقدامات دفاعی استفاده شوند.
- **جهش کد:**
  - از هوش مصنوعی برای تغییر تصادفی کد بدافزار استفاده می‌کند در حالی که عملکرد آن را حفظ می‌کند و انواع چندشکلی یا متامورفیک ایجاد می‌کند.
- **تقلید رفتاری:**
  - از هوش مصنوعی برای شبیه‌سازی رفتار نرم‌افزارهای قانونی استفاده می‌کند و تشخیص بدافزار توسط ابزارهای تحلیل رفتاری را دشوارتر می‌کند.
- **رمزنگاری و فشرده‌سازی:**
  - از هوش مصنوعی برای رمزنگاری یا فشرده‌سازی payloadهای بدافزار استفاده می‌کند تا اطمینان حاصل شود که تا زمان اجرا پنهان می‌مانند.

۲-۱۲-۳ دارایی‌های هدف

- **دستگاه‌های نقطه پایانی:**
  - کامپیوترها، تلفن‌های هوشمند، تبلت‌ها و دستگاه‌های اینترنت اشیا (IoT) که می‌توانند برای

سرقت داده یا اختلال در عملیات مورد هدف قرار گیرند.

- **سیستم‌های شبکه:**
  - شبکه‌های سازمانی، پلتفرم‌های ابری یا سیستم‌های کنترل صنعتی که در برابر انتشار بدافزارهای پنهان‌کار آسیب‌پذیر هستند.
- **راه‌حل‌های ضدویروس:**
  - نرم‌افزارهای ضدویروس سنتی که به تشخیص مبتنی بر امضا یا تحلیل ایستا متکی هستند و می‌توانند توسط مبهم‌سازی مبتنی بر هوش مصنوعی دور زده شوند.
- **محیطهای sandbox:**
  - محیطهای ایزوله امنیتی که برای تحلیل فایل‌های مشکوک استفاده می‌شوند و می‌توانند توسط بدافزارهای تولیدشده توسط هوش مصنوعی که در طول آزمایش بی‌خطر رفتار می‌کنند، فریب داده شوند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تأسیسات بهداشتی، مؤسسات مالی یا سیستم‌های حمل‌ونقل که بدافزارهای کشف‌نشده می‌توانند آسیب‌های قابل توجهی ایجاد کنند.

۴-۱۲-۲ آسیب‌پذیری‌ها

- **تشخیص مبتنی بر امضا:**
  - راه‌حل‌های ضدویروس که از امضاهای قدیمی یا ناقص استفاده می‌کنند، به راحتی توسط بدافزارهای چندشکلی یا متامورفیک دور زده می‌شوند.
- **محدودیت‌های تحلیل ایستا:**
  - عدم وجود قابلیت‌های تحلیل پویا، تشخیص بدافزار مبتنی بر هوش مصنوعی را دشوار می‌کند که تنها در زمان اجرا ماهیت واقعی خود را نشان می‌دهد.
- **شکاف‌های نظارت رفتاری:**
  - ابزارهای تحلیل رفتاری ناکافی نمی‌توانند انحرافات ظریف ناشی از بدافزارهای مبهم‌سازی‌شده توسط هوش مصنوعی را تشخیص دهند.
- **خطای انسانی:**
  - سیاست‌های امنیتی نادرست پیکربندی‌شده یا آموزش ناکافی، سیستم‌ها را در برابر تکنیک‌های مبهم‌سازی پیچیده آسیب‌پذیر می‌کند.

- **آسیب‌پذیری‌های روز صفر:**
  - نقص‌های کشف‌نشده در نرم‌افزار یا سخت‌افزار، نقاط ورودی را برای بدافزار مبتنی بر هوش مصنوعی فراهم می‌کنند تا جای پای خود را محکم کنند.

۲-۱۲-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره سیستم‌های هدف، از جمله دفاع‌ها، پیکربندی‌ها و رفتارهای معمول استفاده می‌کند.
- **طراحی مبهم‌سازی:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل نمونه‌های بدافزار موجود مستقر می‌کند و انواع جدیدی با امضاها، منطق یا رفتار تغییر یافته طراحی می‌کند.
- **تولید payload:**
  - مهاجم بدافزارهای چندشکلی یا متامورفیک ایجاد می‌کند که قادر به فرار از تشخیص هستند در حالی که عملکرد مورد نظر خود را حفظ می‌کنند.
- **آزمایش و بهبود:**
  - مهاجم بدافزار مبهم‌سازی‌شده را در محیط‌های شبیه‌سازی‌شده آزمایش می‌کند تا اطمینان حاصل شود که از اقدامات امنیتی عبور می‌کند و رفتار آن را بر این اساس بهبود می‌بخشد.
- **استقرار:**
  - مهاجم بدافزار مبهم‌سازی‌شده را از طریق ایمیل‌های فیشینگ، کیت‌های سوءاستفاده یا سایر بردارها به سیستم هدف تحویل می‌دهد.
- **پایداری:**
  - مهاجم دسترسی به سیستم‌های به‌خطر افتاده را با تکامل مداوم بدافزار برای سازگاری با تغییرات در دفاع‌ها حفظ می‌کند.

۲-۱۲-۶ نشانه‌های احتمال نفوذ

- **تغییرات غیرعادی فایل:**
  - ناهنجاری‌ها در الگوهای ایجاد، حذف یا تغییر فایل‌ها که با فعالیت‌های قانونی مرتبط نیستند.
- **افزایش استفاده از منابع:**
  - فعالیت بالاتر CPU، حافظه یا I/O دیسک به دلیل اجرای بدافزار مبهم‌سازی‌شده در پس‌زمینه.

- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول سیستم ناسازگار است، مانند اتصالات شبکه غیرمنتظره یا افزایش امتیاز غیرمجاز.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به احراز هویت ناموفق، فرآیندهای غیرمعمول یا فعالیت‌های غیرمنتظره.
- **ترافیک شبکه غیرمنتظره:**
  - انحرافات در ترافیک خروجی، مانند اتصالات به آدرس‌های پروتکل اینترنت IP یا دامنه‌های ناشناخته.

۲-۱۲-۷ تأثیرات

- **نشت داده:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به خسارات مالی یا آسیب به اعتبار می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل به‌خطر افتادن اعتبارنامه‌ها یا سوءاستفاده از آسیب‌پذیری‌ها.
- **خسارات مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلال در خدمات که باعث آسیب مالی می‌شود.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و اقدامات قانونی می‌شود.
- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا فرم‌ورهای حیاتی که باعث غیرقابل استفاده شدن سیستم‌ها می‌شود.

۲-۱۲-۸ راه‌های مقابله

- **هوشمندسازی تهدیدات پیشرفته:**
  - از پلتفرم‌های هوشمندسازی تهدیدات مبتنی بر هوش مصنوعی استفاده کنید تا درباره

تکنیک‌های مبهم‌سازی نوظهور و بدافزارهای سازگار به‌روز بمانید.

- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص فعالیت‌های غیرعادی که نشان‌دهنده بدافزار مبهم‌سازی‌شده هستند، استفاده کنید حتی اگر امضاهای شناخته‌شده نداشته باشد.
- **تحلیل پویا:**
  - محیط‌های sandbox را پیاده‌سازی کنید که قادر به اجرا و تحلیل فایل‌های مشکوک در تنظیمات کنترل‌شده هستند تا رفتارهای پنهان را آشکار کنند.
- **تشخیص مبتنی بر یادگیری ماشین:**
  - راه‌حل‌های ضدویروس مبتنی بر یادگیری ماشین را مستقر کنید که بر روی مجموعه‌های بزرگ داده‌های نمونه‌های بدافزار آموزش دیده‌اند تا تهدیدات مبهم‌سازی‌شده را به‌طور مؤثرتری شناسایی کنند.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرم‌ورها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده که توسط بدافزارهای مبهم‌سازی‌شده سوءاستفاده می‌شوند، برطرف شوند.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش بدافزار محدود شود و تأثیر عفونت‌های موفق را کاهش دهید.
- **بهداشت رمزنگاری:**
  - داده‌های حساس را هم در حالت استراحت و هم در حال انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده را به حداقل برسانید.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد تشخیص علائم عفونت بدافزار، مانند تلاش‌های فیشینگ یا رفتار غیرعادی سیستم آموزش دهید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی ایجاد و پیاده‌سازی کنید تا به‌سرعت حملات بدافزار

مبهم‌سازی‌شده را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.

۲-۱۳ تولید داده‌های مصنوعی

تولید داده‌های مصنوعی[1] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای ایجاد مجموعه‌داده‌های واقع‌گرایانه اما مصنوعی است که اطلاعات دنیای واقعی را تقلید می‌کنند. این تکنیک به‌طور فزاینده‌ای در هر دو زمینه قانونی و مخرب استفاده می‌شود. در حوزه امنیت سایبری، مهاجمان می‌توانند از هوش مصنوعی برای تولید داده‌های مصنوعی با اهداف فریب‌آمیز استفاده کنند، مانند دور زدن سیستم‌های تشخیص، آموزش مدل‌های تقابلی یا راه‌اندازی حملات پیشرفته‌تر مانند فیشینگ، کلاهبرداری یا سرقت هویت. از سوی دیگر، مدافعان نیز می‌توانند از تولید داده‌های مصنوعی برای بهبود اقدامات امنیتی، ارتقای مدل‌های یادگیری ماشین و شبیه‌سازی سناریوهای حمله استفاده کنند.

در حالی که تولید داده‌های مصنوعی کاربردهای مثبت بسیاری دارد، سوءاستفاده از آن توسط مهاجمان خطرات قابل توجهی برای سازمان‌ها و افراد ایجاد می‌کند.

۲-۱۳-۱ ویژگی‌های کلیدی

- **واقع‌گرایی بالا:**
  - هوش مصنوعی داده‌های مصنوعی تولید می‌کند که تقریباً غیرقابل تشخیص از داده‌های واقعی هستند و آن را برای فریب یا شبیه‌سازی مؤثر می‌سازد.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند تولید حجم زیادی از داده‌های مصنوعی را خودکار می‌کنند، که تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **انطباق‌پذیری:**
  - تولید داده‌های مصنوعی مبتنی بر هوش مصنوعی می‌تواند با گذشت زمان تکامل یابد و با تغییرات در پیکربندی سیستم‌ها، اقدامات دفاعی یا شرایط محیطی سازگار شود.
- **مقیاس‌پذیری:**
  - هوش مصنوعی امکان تولید سریع مجموعه‌داده‌های عظیم را فراهم می‌کند که متناسب با اهداف خاص، چه برای اهداف قانونی و چه برای اهداف مخرب، تنظیم شده‌اند.
- **سفارشی‌سازی:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا داده‌های مصنوعی را متناسب با اهداف

1. Synthetic data generation

خاص، صنایع یا بردارهای حمله تنظیم کنند.

۲-۱۳-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های برچسب‌دار از داده‌های واقعی آموزش می‌دهد تا الگوها را یاد بگیرد و داده‌های مصنوعی با ویژگی‌های مشابه تولید کند.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در داده‌های بدون برچسب را شناسایی می‌کند تا روابط پنهان را آشکار کند و استراتژی‌های تولید را بهبود بخشد.
  - یادگیری تقویتی: تولید داده‌های مصنوعی را با پاداش دادن به نتایج موفق و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های مولد تخاصمی، داده‌های مصنوعی اما واقع‌گرایانه ایجاد می‌کنند که توزیع‌های پیچیده موجود در مجموعه‌داده‌های واقعی را شبیه‌سازی می‌کنند.
  - ترنسفورمرها: داده‌های ترتیبی، مانند متن یا اطلاعات سری‌زمانی، را تحلیل می‌کنند تا خروجی‌های مصنوعی منسجم و مرتبط با زمینه تولید کنند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی مصنوعی، مانند ایمیل‌های جعلی، پست‌های شبکه‌های اجتماعی یا پیام‌های چت، ایجاد می‌کند که سبک‌های ارتباطی انسانی را تقلید می‌کنند.
- **بینایی کامپیوتر:**
  - تصاویر، ویدیوها یا داده‌های بیومتریک مصنوعی (مانند چهره‌ها، اثر انگشت) ایجاد می‌کند که برای جعل هویت یا تکنیک‌های دور زدن استفاده می‌شوند.
- **افزایش داده:**
  - از هوش مصنوعی برای گسترش مجموعه‌داده‌های موجود با ایجاد تغییراتی در داده‌های واقعی استفاده می‌کند تا استحکام مدل‌ها را بهبود بخشد یا اثربخشی حمله را افزایش دهد.

۲-۱۳-۳ دارایی‌های هدف

- **سیستم‌های تشخیص:**
  - مهاجمان داده‌های مصنوعی تولید می‌کنند تا سیستم‌های تشخیص مانند نرم‌افزارهای آنتی‌ویروس، سیستم‌های تشخیص نفوذ (IDS) یا ابزارهای تحلیل رفتاری را دور بزنند.

- **اطلاعات حساس:**
  - ایمیل‌ها، پیام‌ها یا وب‌سایت‌های مصنوعی ایجاد می‌شوند تا کاربران را فریب دهند و اطلاعات حساس را افشا کنند.
- **فعالیت‌های کلاهبرداری:**
  - سوابق تراکنش‌ها، هویت‌ها یا اعتبارنامه‌های جعلی ایجاد می‌شوند تا کلاهبرداری مالی انجام شود یا منابع سرقت شوند.
- **کمپین‌های اطلاعات نادرست:**
  - محتوای چندرسانه‌ای مصنوعی، مانند جعل عمیق‌ها یا تصاویر دستکاری‌شده، ایجاد می‌شود تا اطلاعات نادرست را در مقیاس گسترده منتشر کند.
- **شبیه‌سازی تهدیدات داخلی:**
  - مهاجمان رفتار داخلی را با استفاده از داده‌های مصنوعی شبیه‌سازی می‌کنند تا سیستم‌های تشخیص ناهنجاری را دور بزنند.

۴-۱۳-۲ آسیب‌پذیری‌ها

- **مکانیزم‌های تشخیص ضعیف:**
  - سیستم‌های تشخیص سنتی ممکن است نتوانند بین داده‌های مصنوعی و واقعی تمایز قائل شوند، که به مهاجمان امکان می‌دهد از این شکاف سوءاستفاده کنند.
- **اتکای بیش از حد به خودکارسازی:**
  - سیستم‌هایی که تنها به ابزارهای خودکار برای اعتبارسنجی یا طبقه‌بندی متکی هستند، ممکن است توسط داده‌های مصنوعی بسیار واقع‌گرایانه فریب بخورند.
- **خطای انسانی:**
  - آموزش یا آگاهی ناکافی در میان کارمندان می‌تواند منجر به پذیرش داده‌های مصنوعی به‌عنوان داده‌های قانونی شود.
- **دفاع‌های قدیمی:**
  - استفاده از دفاع‌های قدیمی یا نادرست پیاده‌سازی‌شده، سازمان‌ها را در برابر حملات تطبیقی داده‌های مصنوعی آسیب‌پذیر می‌کند.
- **فقدان اقدامات پیش‌گیرانه:**
  - فقدان اقدامات پیش‌گیرانه برای تشخیص یا کاهش تهدیدات مبتنی بر داده‌های مصنوعی،

خطر سوءاستفاده را افزایش می‌دهد.

۲-۱۳-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری داده‌های دنیای واقعی که به‌عنوان پایه‌ای برای تولید داده‌های مصنوعی استفاده می‌شود، استفاده می‌کند.
- **آموزش مدل:**
  - مهاجم مدل‌های یادگیری ماشین، مانند شبکه‌های مولد تخاصمی، را روی داده‌های جمع‌آوری‌شده آموزش داده تا الگوها را یاد بگیرند و داده‌های مصنوعی متناظر ایجاد کنند.
- **ایجاد داده‌های مصنوعی:**
  - مهاجم داده‌های مصنوعی متناسب با اهداف خاص، مانند ایجاد هویت‌ها، تراکنش‌ها یا ارتباطات جعلی، ایجاد می‌کند.
- **استقرار:**
  - مهاجم داده‌های مصنوعی را در محیط هدف مستقر می‌کند و اطمینان حاصل می‌کند که به‌طور یکپارچه با داده‌های واقعی ترکیب می‌شود.
- **سوءاستفاده:**
  - مهاجم از داده‌های مصنوعی برای اهداف مخرب، مانند دور زدن تشخیص، انجام کلاهبرداری یا انتشار اطلاعات نادرست، استفاده می‌کند.
- **پایداری:**
  - مهاجم داده‌های مصنوعی را بر اساس بازخورد به‌طور مداوم بهبود می‌بخشد تا اثربخشی بلندمدت حفظ شود.

۲-۱۳-۶ نشانه‌های احتمال نفوذ

- **الگوهای غیرعادی داده:**
  - همبستگی‌ها یا ناسازگاری‌های غیرمعمول در مجموعه‌داده‌ها که از رفتارهای مورد انتظار منحرف می‌شوند.
- **افزایش حجم داده:**
  - افزایش ناگهانی در حجم یا فرکانس داده‌ها که با عملیات قانونی مرتبط نیست.

- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر یا سیستم ناسازگار هستند، مانند الگوهای دسترسی غیرمنتظره یا پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره یا فعالیت‌های غیرمجاز.
- **نتایج غیرمنتظره:**
  - انحرافات در خروجی‌های سیستم، پیش‌بینی‌ها یا تصمیم‌گیری‌ها که ممکن است نشان‌دهنده تأثیر داده‌های مصنوعی باشد.

۲-۱۳-۷ تأثیرات

- **اختلال در عملیات:**
  - داده‌های مصنوعی می‌توانند فرآیندهای کسب‌وکار، سیستم‌های تصمیم‌گیری یا زیرساخت‌های حیاتی را با معرفی ورودی‌های گمراه‌کننده مختل کنند.
- **ضررهای مالی:**
  - فعالیت‌های کلاهبرداری که توسط داده‌های مصنوعی امکان‌پذیر می‌شوند، مانند تراکنش‌ها یا هویت‌های جعلی، منجر به آسیب‌های مالی می‌شوند.
- **آسیب به شهرت:**
  - سازمان‌ها ممکن است در صورت استفاده از داده‌های مصنوعی برای دستکاری افکار عمومی یا تضعیف اعتماد، آسیب اعتباری ببینند.
- **جریمه‌های نظارتی:**
  - عدم رعایت مقررات حفاظت از داده‌ها به دلیل سوءاستفاده یا افشای داده‌های مصنوعی.
- **فساد سیستم:**
  - داده‌های مصنوعی می‌توانند پایگاه‌داده‌ها، مدل‌ها یا لاگ‌ها را فاسد کنند و سیستم‌ها را غیرقابل اعتماد یا غیرقابل استفاده کنند.

۲-۱۳-۸ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به شناسایی تفاوت‌های

ظریف بین داده‌های مصنوعی و واقعی از طریق تشخیص الگوهای پیشرفته هستند.

- **تحلیل رفتاری:**
  - از تحلیل رفتاری مبتنی بر هوش مصنوعی استفاده کنید تا الگوهای فعالیت غیرعادی که نشان‌دهنده استفاده از داده‌های مصنوعی هستند، شناسایی شوند.
- **بررسی‌های یکپارچگی داده:**
  - مکانیزم‌هایی مانند هش‌کردن، امضای دیجیتال یا تأیید مبتنی بر بلاک‌چین را پیاده‌سازی کنید تا اصالت داده‌ها را تضمین کنید.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های مکرر از مجموعه‌داده‌ها، لاگ‌ها و خروجی‌های سیستم انجام دهید تا مسائل بالقوه مربوط به داده‌های مصنوعی را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **رمزنگاری و کنترل‌های دسترسی:**
  - داده‌های حساس را رمزنگاری کنید و کنترل‌های دسترسی سخت‌گیرانه اعمال کنید تا تغییرات غیرمجاز یا معرفی داده‌های مصنوعی محدود شود.
- **ردیابی منشأ:**
  - سوابق دقیقی از منشأ داده‌ها و تغییرات آن‌ها نگه دارید تا منشأ و اعتبار تمام ورودی‌ها را ردیابی و تأیید کنید.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد شناسایی نشانه‌های سوءاستفاده از داده‌های مصنوعی و رعایت بهداشت سایبری آموزش دهید.
- **هوشمندسازی تهدیدات پیشرفته:**
  - از پلتفرم‌های هوشمند مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور داده‌های مصنوعی و تهدیدات تطبیقی مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی‌ای توسعه دهید و پیاده‌سازی کنید تا حملات داده‌های مصنوعی را به‌سرعت شناسایی، مهار و کاهش دهید.
- **اعتبارسنجی داده‌های مصنوعی:**
  - از ابزارهای تخصصی برای اعتبارسنجی داده‌های مصنوعی قبل از ادغام آن‌ها در سیستم‌ها یا

گردش کارهای حیاتی استفاده کنید.

۲-۱۴ تولید هوشمند نظرات جعلی

تولید هوشمند نظرات جعلی[1] شامل استفاده از هوش مصنوعی برای ایجاد و انتشار خودکار نظرات بسیار واقع‌گرایانه اما جعلی برای محصولات، خدمات یا کسب‌وکارها است. با بهره‌گیری از یادگیری ماشین و پردازش زبان طبیعی، مهاجمان می‌توانند نظراتی ایجاد کنند که سبک نوشتاری، احساسات و نظرات انسانی را تقلید می‌کنند و تشخیص آن‌ها را دشوارتر می‌سازند. این نوع حمله به‌خصوص خطرناک است زیرا اعتماد به پلتفرم‌های آنلاین را تضعیف می‌کند، رفتار مصرف‌کنندگان را دستکاری می‌کند و می‌تواند آسیب‌های جدی به اعتبار کسب‌وکارها یا افراد وارد کند.

تولید هوشمند نظرات جعلی مبتنی بر هوش مصنوعی به مهاجمان امکان می‌دهد عملیات خود را به‌طور کارآمد مقیاس‌پذیر کنند و چندین پلتفرم یا رقبا را به‌طور همزمان هدف قرار دهند. همچنین امکان تنظیمات پویا بر اساس بازخورد بلادرنگ را فراهم می‌کند و اثربخشی بلندمدت را تضمین می‌کند.

۲-۱۴-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند تولید، ارسال و مدیریت نظرات جعلی را خودکار می‌کند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **واقع‌گرایی بالا:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند نظراتی ایجاد کنند که تقریباً از نظرات واقعی غیرقابل تشخیص هستند و اعتبار را افزایش می‌دهند.
- **انعطاف‌پذیری:**
  - سیستم‌های مبتنی بر هوش مصنوعی به‌طور پویا تکامل می‌یابند و با تغییرات در سیاست‌های پلتفرم، مکانیسم‌های تشخیص یا شرایط بازار سازگار می‌شوند.
- **هدف‌گیری دقیق:**
  - مهاجمان می‌توانند بر رقبای خاص، صنایع یا زمینه‌ها تمرکز کنند و نظرات را برای حداکثر تأثیر سفارشی‌سازی کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی امکان تولید و توزیع سریع حجم زیادی از نظرات جعلی در چندین پلتفرم

1. Smart Fake Review Generation

را فراهم می‌کند.

۲-۱۴-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **پردازش زبان طبیعی:**
  - تولید متن: از مدل‌های مبتنی بر ترنسفورمر مانند GPT-3 یا BERT برای ایجاد نظرات منسجم و مرتبط با زمینه استفاده می‌کند.
  - تحلیل احساسات: احساسات و تن‌های عاطفی را تحلیل می‌کند تا نظراتی ایجاد کند که با ترجیحات مخاطبان هدف همسو باشد.
  - مدل‌سازی موضوعی: موضوعات یا تم‌های پرطرفدار را شناسایی می‌کند تا اطمینان حاصل کند که نظرات با علایق یا رویدادهای فعلی همخوانی دارند.
- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های برچسب‌دار از نظرات جعلی موفق و ناموفق آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا خوشه‌ها در نظرات واقعی را شناسایی می‌کند تا نشانه‌های سبکی را کشف کرده و واقع‌گرایی را بهبود بخشد.
  - یادگیری تقویتی: تولید نظرات را با پاداش‌دهی به معیارهای مشارکت بالا (مانند لایک‌ها، اشتراک‌گذاری‌ها) و جریمه‌کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، الگوهای پیچیده در محتوای تولیدشده توسط کاربر را تحلیل می‌کنند تا لحن، ساختار و قانع‌کنندگی را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: نظرات مصنوعی اما واقع‌گرایانه ایجاد می‌کنند که برای دور زدن سیستم‌های تشخیص طراحی شده‌اند.
- **تقلید رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر استفاده می‌کند تا الگوهای ارسال قانونی را شبیه‌سازی کند.
- **داده‌کاوی:**
  - داده‌های عمومی را از شبکه‌های اجتماعی، فروم‌ها یا وب‌سایت‌های رقبا استخراج می‌کند تا پروفایل‌های دقیقی از مخاطبان یا صنایع هدف ایجاد کند.

۲-۱۴-۳ دارایی‌های هدف

- **پلتفرم‌های تجارت الکترونیک:**
  - بازارهای آنلاین که در آن‌ها نظرات جعلی می‌توانند بر تصمیمات خرید تأثیر بگذارند یا به اعتبار رقبا آسیب برسانند.
- **ارائه‌دهندگان خدمات:**
  - کسب‌وکارهایی مانند رستوران‌ها، هتل‌ها یا مشاوران که اعتبار برای موفقیت آن‌ها حیاتی است.
- **تولیدکنندگان محصولات:**
  - شرکت‌هایی که کالاهایی تولید می‌کنند و برای فروش و وفاداری برند به بازخورد مشتریان وابسته هستند.
- **وب‌سایت‌های بررسی:**
  - پلتفرم‌هایی مانند TripAdvisor، Yelp یا Amazon که در آن‌ها نظرات جعلی می‌توانند رتبه‌بندی‌ها یا ادراکات را دستکاری کنند.
- **رقبا:**
  - رقبای بازار که ممکن است از نظرات جعلی برای تضعیف اعتبار یکدیگر یا بهبود رتبه‌بندی خود استفاده کنند.

۲-۱۴-۴ آسیب‌پذیری‌ها

- **مکانیسم‌های تشخیص ضعیف:**
  - پلتفرم‌هایی که برای تشخیص نظرات جعلی به سیستم‌های قدیمی یا مبتنی بر قوانین متکی هستند، ممکن است نتوانند محتوای تولیدشده توسط هوش مصنوعی را شناسایی کنند.
- **روانشناسی انسانی:**
  - مصرف‌کنندگان بیشتر احتمال دارد به نظراتی اعتماد کنند که واقعی به نظر می‌رسند و در نتیجه در برابر دستکاری آسیب‌پذیر هستند.
- **اتکای بیش از حد به خودکارسازی:**
  - ابزارهای مدیریت خودکار ممکن است در تشخیص تفاوت بین نظرات واقعی و مصنوعی به دلیل سطح بالای واقع‌گرایی آن‌ها مشکل داشته باشند.

- **عدم وجود اقدامات پیشگیرانه:**
  - عدم وجود اقدامات قوی برای تشخیص و حذف نظرات جعلی، خطر سوءاستفاده را افزایش می‌دهد.
- **داده‌های عمومی:**
  - مجموعه‌داده‌های نشت‌یافته یا استخراج‌شده حاوی نظرات واقعی، مواد ارزشمندی برای آموزش مدل‌های هوش مصنوعی جهت ایجاد نظرات جعلی متقاعدکننده فراهم می‌کنند.

۵-۱۴-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری مجموعه‌داده‌های بزرگ از نظرات واقعی از پلتفرم‌ها یا صنایع هدف استفاده می‌کند.
- **آموزش مدل:**
  - مهاجم مدل‌های یادگیری ماشین، مانند GPT-3 یا BERT، را روی داده‌های جمع‌آوری‌شده آموزش می‌دهد تا الگوها، احساسات و ساختار نظرات واقعی را یاد بگیرد.
- **تولید نظر:**
  - مهاجم مدل‌های آموزش‌دیده را برای ایجاد نظرات جعلی سفارشی‌سازی‌شده برای اهداف، صنایع یا اهداف خاص به‌کار می‌گیرد.
- **استقرار:**
  - مهاجم ارسال خودکار نظرات تولیدشده را از طریق چندین حساب یا پلتفرم برای دستیابی به نتایج مطلوب انجام می‌دهد.
- **تقویت:**
  - مهاجم از توصیه‌های الگوریتمی، موضوعات پرطرفدار یا شبکه‌های تأثیرگذار بهره‌برداری می‌کند تا حداکثر دسترسی و تأثیر را ایجاد کند.
- **پایداری:**
  - مهاجم کمپین را بر اساس بازخورد به‌طور مداوم بهبود می‌بخشد و تاکتیک‌ها را تشدید یا محتوا را تغییر می‌دهد تا اثربخشی را حفظ کند.

۲-۱۴-۶ نشانه‌های احتمال نفوذ

- **الگوهای ارسال غیرعادی:**
  - ناهنجاری‌ها در فرکانس ارسال، زمان‌بندی یا فعالیت حساب که با کاربران قانونی مرتبط نیستند.
- **محتواهای مشابه:**
  - نظرات با عبارات، ساختارها یا احساسات تکراری که از تنوع معمول در محتوای تولیدشده توسط کاربر منحرف می‌شوند.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار عادی کاربر، مانند ایجاد چندین حساب با الگوهای ارسال مشابه.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های پلتفرم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.
- **نتایج غیرمنتظره:**
  - تغییرات ناگهانی در رتبه‌بندی محصولات، بررسی خدمات یا رتبه‌بندی‌های جستجو که ممکن است نشان‌دهنده تأثیر نظرات جعلی باشد.

۲-۱۴-۷ تأثیرات

- **آسیب به اعتبار:**
  - قربانیان نظرات جعلی آسیب‌های اعتباری می‌بینند که منجر به از دست دادن مشتریان، شرکا یا سهم بازار می‌شود.
- **ضررهای مالی:**
  - نظرات دستکاری‌شده می‌توانند منجر به کاهش فروش، درآمد کمتر یا افزایش هزینه‌ها برای کنترل آسیب شوند.
- **تحریف بازار:**
  - نظرات جعلی رقابت را تحریف می‌کنند و به برخی کسب‌وکارها یا محصولات مزایای ناعادلانه می‌دهند در حالی که به دیگران آسیب می‌رسانند.
- **فریب مصرف‌کننده:**
  - اطلاعات گمراه‌کننده باعث می‌شود مصرف‌کنندگان تصمیمات خرید ضعیفی بگیرند و اعتماد

به پلتفرم‌های آنلاین را تضعیف کنند.

- **فرسایش پلتفرم:**
  - با گذشت زمان، پلتفرم‌هایی که با نظرات جعلی مواجه می‌شوند اعتبار خود را از دست می‌دهند و بر مشارکت کاربر و سودآوری تأثیر می‌گذارند.

۸-۱۴-۲ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به شناسایی آثار ظریف یا ناسازگاری‌ها در نظرات مصنوعی هستند.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت که نشان‌دهنده کمپین‌های نظرات جعلی است، استفاده کنید.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های مکرر از نظرات و حساب‌های کاربری انجام دهید تا فعالیت‌های مشکوک را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **احراز هویت کاربر:**
  - فرآیندهای احراز هویت قوی کاربر، مانند احراز هویت چندعاملی یا بررسی هویت، را پیاده‌سازی کنید تا ایجاد حساب‌های جعلی کاهش یابد.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور نظرات جعلی و تهدیدات انطباق‌پذیر مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت کمپین‌های نظرات جعلی را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.
- **دفاع پیشگیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های در حال تکامل نظرات جعلی استفاده کنید.

- **احراز هویت محتوا:**
  - مکانیسم‌های تأیید مبتنی بر بلاک‌چین یا رمزنگاری را پیاده‌سازی کنید تا اصالت محتوای تولیدشده توسط کاربر را تضمین کنند.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد شناسایی علائم نظرات جعلی آموزش دهید و فرهنگ هوشیاری را در بین مصرف‌کنندگان و مدیران تقویت کنید.
- **تلاش‌های مشارکتی:**
  - همکاری بین پلتفرم‌ها، دولت‌ها و محققان را تقویت کنید تا استانداردها و راه‌حل‌های مشترک برای مقابله با نظرات جعلی توسعه یابد.

۲-۱۵ حمله به CAPTCHA

CAPTCHA (آزمون تورینگ عمومی کاملاً خودکار برای تشخیص انسان و رایانه) یک مکانیزم امنیتی است که برای تشخیص کاربران انسانی از سیستم‌های خودکار طراحی شده است و از کاربران می‌خواهد چالش‌هایی مانند شناسایی متن تحریف‌شده، کلیک بر روی تصاویر خاص یا حل پازل‌های منطقی را حل کنند. در حمله به CAPTCHA مبتنی بر هوش مصنوعی، از هوش مصنوعی برای خودکارسازی فرآیند حل CAPTCHA با دقت بالا استفاده می‌شود، که به مهاجمان امکان می‌دهد از این دفاع‌ها عبور کرده و فعالیت‌های مخربی مانند اسپم، حملات سرقت هویت یا تصاحب حساب‌ها را انجام دهند.

با استفاده از تکنیک‌های یادگیری ماشین و بینایی کامپیوتری، حملات CAPTCHA مبتنی بر هوش مصنوعی می‌توانند چالش‌های CAPTCHA را به‌طور مؤثرتری نسبت به روش‌های سنتی بروت فورس تحلیل و رمزگشایی کنند و بسیاری از سیستم‌های CAPTCHA را منسوخ کنند.

۲-۱۵-۱ ویژگی‌های کلیدی

- **خودکارسازی:**

هوش مصنوعی فرآیند حل CAPTCHA را در مقیاس بزرگ خودکار می‌کند، که باعث کاهش تلاش دستی و افزایش کارایی می‌شود.

- **دقت بالا:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند حتی چالش‌های پیچیده CAPTCHA را با دقت نزدیک به انسان حل کنند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در

طراحی‌ها، الگوریتم‌ها یا شرایط محیطی CAPTCHA سازگار شوند.

- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد هزاران یا میلیون‌ها CAPTCHA را به‌طور همزمان حل کنند، که دامنه کمپین‌های آن‌ها را افزایش می‌دهد.
- **پنهان‌کاری:**
  - با تقلید از رفتار کاربران قانونی در تعاملات CAPTCHA، هوش مصنوعی اطمینان می‌دهد که فعالیت‌های مخرب بدون تشخیص باقی بمانند.

۲-۱۵-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های برچسب‌خورده از CAPTCHAهای حل‌شده آموزش می‌دهد تا الگوها را یاد بگیرد و دقت را بهبود بخشد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در طراحی‌های CAPTCHA شناسایی می‌کند تا نقاط ضعف بالقوه را بدون نیاز به برچسب‌گذاری قبلی آشکار کند.
  - یادگیری تقویتی: استراتژی‌های حل CAPTCHA را با پاداش دادن به نتایج موفق و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی، عناصر بصری CAPTCHA (مانند متن تحریف‌شده یا تصاویر) را تحلیل می‌کنند تا الگوها را تشخیص دهند و چالش‌ها را رمزگشایی کنند.
  - شبکه‌های عصبی بازگشتی انواع CAPTCHAهای متوالی، مانند CAPTCHAهای صوتی یا پازل‌های منطقی، را مدیریت می‌کنند.
- **بینایی کامپیوتری:**
  - از تکنیک‌های پیشرفته تشخیص تصویر برای شناسایی اشیاء، متن یا الگوها در CAPTCHAهای بصری استفاده می‌کند، که امکان حل خودکار را فراهم می‌کند.
- **پردازش زبان طبیعی:**
  - چالش‌های متنی یا مفهومی CAPTCHA، مانند پازل‌های منطقی یا تکمیل جملات، را تحلیل می‌کند تا پاسخ‌های صحیح تولید کند.

- **افزایش داده:**
  - مجموعه‌داده‌های آموزشی را با تولید تغییراتی در چالش‌های CAPTCHA گسترش می‌دهد، که استحکام و انعطاف‌پذیری مدل را بهبود می‌بخشد.

۲-۱۵-۳ دارایی‌های هدف

- **برنامه‌های وب:**
  - فرم‌های ورود به سیستم، صفحات ثبت‌نام یا فرم‌های تماس که توسط مکانیزم‌های CAPTCHA محافظت می‌شوند.
- **خدمات آنلاین:**
  - ارائه‌دهندگان ایمیل، پلتفرم‌های شبکه‌های اجتماعی یا سایت‌های تجارت الکترونیک که دسترسی خودکار می‌تواند منجر به سوءاستفاده یا کلاهبرداری شود.
- **سیستم‌های مالی:**
  - پورتال‌های بانکی، درگاه‌های پرداخت یا صرافی‌های ارز دیجیتال که در برابر حملات خودکار آسیب‌پذیر هستند.
- **وب‌سایت‌های دولتی:**
  - اهداف باارزش که اطلاعات حساس را ذخیره می‌کنند یا خدمات حیاتی ارائه می‌دهند، جایی که دور زدن CAPTCHA می‌تواند منجر به نقض داده یا اختلال در خدمات شود.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - رابط‌های وب برای دستگاه‌های اینترنت اشیا که اغلب توسط مکانیزم‌های CAPTCHA ساده محافظت می‌شوند و به راحتی توسط هوش مصنوعی دور زده می‌شوند.

۲-۱۵-۴ آسیب‌پذیری‌ها

- **طراحی‌های ضعیف CAPTCHA:**
  - CAPTCHAهای دارای الگوهای قابل پیش‌بینی، پیچیدگی ناکافی یا الگوریتم‌های قدیمی به راحتی توسط هوش مصنوعی حل می‌شوند.
- **اعتبارسنجی ناکافی:**
  - نبود مکانیزم‌های اعتبارسنجی قوی در پشت صحنه یا محدودیت‌های نرخ، به مهاجمان امکان می‌دهد به‌طور مکرر تلاش‌هایی برای حل CAPTCHA انجام دهند.

- **خطای انسانی:**
  - توسعه‌دهندگانی که فناوری‌های مدرن CAPTCHA را پیاده‌سازی نمی‌کنند یا به‌روزرسانی‌ها را نادیده می‌گیرند، سیستم‌ها را در برابر حملات مبتنی بر هوش مصنوعی آسیب‌پذیر می‌کنند.
- **تکیه بیش از حد به چالش‌های ایستا:**
  - CAPTCHAهایی که به چالش‌های ایستا یا تکراری متکی هستند، زمینه مناسبی برای آموزش و سوءاستفاده هوش مصنوعی فراهم می‌کنند.
- **عدم نظارت رفتاری:**
  - نبود ابزارهایی برای تشخیص الگوهای غیرمعمول تعامل با CAPTCHA، سیستم‌ها را در برابر حملات خودکار آسیب‌پذیر می‌کند.

۵-۱۵-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری مجموعه‌داده‌های بزرگ از CAPTCHAهای حل‌شده یا نمونه‌های CAPTCHA در دسترس عموم برای اهداف آموزشی استفاده می‌کند.
- **آموزش مدل:**
  - مهاجم مدل‌های یادگیری ماشین، به‌ویژه شبکه عصبی پیچشی برای CAPTCHAهای مبتنی بر تصویر یا شبکه‌های عصبی بازگشتی برای CAPTCHAهای متنی/صوتی، را آموزش داده تا الگوها را تشخیص دهند و دقت را بهبود بخشند.
- **حل CAPTCHA:**
  - مهاجم مدل آموزش‌دیده را برای حل خودکار چالش‌های CAPTCHA در زمان واقعی مستقر کرده تا دسترسی بی‌وقفه به سیستم‌های هدف فراهم شود.
- **استقرار:**
  - مهاجم تعاملات با برنامه‌های وب یا خدمات را خودکار می‌کند و از محافظت‌های CAPTCHA عبور کرده تا فعالیت‌های مخربی مانند اسپم، scraping یا حملات credential stuffing انجام دهید.
- **پایداری:**
  - مهاجم مدل را به‌طور مداوم بر اساس بازخورد بهبود می‌بخشد تا اثربخشی آن در برابر

طراحی‌های در حال تحول CAPTCHA حفظ شود.

۲-۱۵-۶ نشانه‌های احتمال نفوذ

- **افزایش تلاش‌های CAPTCHA:**
  - افزایش ناگهانی در تلاش‌های حل CAPTCHA از یک آدرس پروتکل اینترنت (IP) یا عامل کاربر واحد.
- **نرخ موفقیت غیرمعمول:**
  - نرخ موفقیت غیرعادی بالا در حل چالش‌های CAPTCHA، که نشان‌دهنده استفاده از تکنیک‌های پیشرفته هوش مصنوعی است.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر ناسازگار هستند، مانند تعاملات سریع یا دسترسی به مناطق محدود بلافاصله پس از دور زدن CAPTCHA.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به تعاملات مکرر CAPTCHA یا فعالیت‌های غیرمنتظره.
- **الگوهای ترافیکی غیرمنتظره:**
  - انحراف در ترافیک شبکه، مانند اتصالات از بات‌نت‌ها یا حساب‌های نفوذشده که سعی در دور زدن محافظت‌های CAPTCHA دارند.

۲-۱۵-۷ تأثیرات

- **اسپم و سوءاستفاده:**
  - دور زدن CAPTCHAها به مهاجمان امکان می‌دهد سیستم‌ها را با اسپم، حساب‌های جعلی یا محتوای مخرب بمباران کنند.
- **حمله به اعتبارنامه‌ها:**
  - دسترسی خودکار به فرم‌های ورود به سیستم به مهاجمان امکان می‌دهد اعتبارهای سرقت‌شده را در مقیاس بزرگ آزمایش کنند، که منجر به تصاحب حساب‌ها می‌شود.
- **استخراج داده:**
  - دسترسی غیرمجاز به پایگاه‌داده‌ها یا APIها منجر به سرقت داده، از دست دادن مالکیت فکری یا مزیت‌های رقابتی می‌شود.

- **اختلال عملیاتی:**
  - بمباران سیستم‌ها با درخواست‌های خودکار، عملیات عادی را مختل می‌کند و باعث خرابی یا کاهش عملکرد می‌شود.
- **آسیب به شهرت:**
  - دور زدن موفقیت‌آمیز CAPTCHAها اعتبار سازمان‌هایی را که ناامن یا غیرقابل اعتماد تلقی می‌شوند، خدشه‌دار می‌کند.

۸-۱۵-۲ راه‌های مقابله

- **فناوری‌های پیشرفته CAPTCHA:**
  - سیستم‌های CAPTCHA مدرن، مانند Google reCAPTCHA v3، را پیاده‌سازی کنید که به جای چالش‌های ایستا، بر تحلیل رفتاری تکیه می‌کنند.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص ناهنجاری‌ها در تعاملات کاربر، مانند زمان‌های غیرعادی حل CAPTCHA یا الگوهای ناوبری ربات‌گونه استفاده کنید.
- **محدودیت نرخ:**
  - محدودیت‌های نرخ سخت‌گیرانه‌ای بر تلاش‌های CAPTCHA اعمال کنید تا از بمباران سیستم‌ها توسط سیستم‌های خودکار جلوگیری شود.
- **چالش‌های پویا:**
  - چالش‌های CAPTCHA را طراحی کنید که به‌طور پویا در پیچیدگی، فرمت یا ارائه تغییر می‌کنند تا حل‌کننده‌های مبتنی بر هوش مصنوعی را خنثی کنند.
- **شناسایی و احراز هویت دستگاه:**
  - از تکنیک‌های شناسایی دستگاه استفاده کنید تا دستگاه‌ها یا مرورگرهای مشکوکی که سعی در دور زدن محافظت‌های CAPTCHA دارند را شناسایی و مسدود کنید.
- **به‌روزرسانی‌های منظم:**
  - سیستم‌های CAPTCHA را با پیشرفت‌های اخیر در طراحی‌ها و الگوریتم‌های مقاوم در برابر هوش مصنوعی به‌روز نگه دارید.
- **آموزش و آگاهی:**
  - توسعه‌دهندگان و مدیران را در مورد پیاده‌سازی مکانیزم‌های CAPTCHA ایمن و تشخیص

علائم حملات خودکار آموزش دهید.

- **دفاع چندلایه:**
  - محافظت‌های CAPTCHA را با سایر اقدامات امنیتی، مانند احراز هویت چندعاملی یا مسدودسازی IP، ترکیب کنید تا وابستگی به CAPTCHA به تنهایی کاهش یابد.
- **هوش تهدید پیش‌گیرانه:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا درباره تکنیک‌های نوظهور حل CAPTCHA و تهدیدات تطبیقی به‌روز بمانید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت تلاش‌های دور زدن CAPTCHA را تشخیص، مهار و کاهش دهید.

۲-۱۶ تولید خودکار اطلاعات نادرست

تولید خودکار اطلاعات نادرست با استفاده از هوش مصنوعی شامل به‌کارگیری فناوری‌های هوش مصنوعی برای ایجاد، انتشار و تقویت اطلاعات نادرست یا گمراه‌کننده در مقیاس وسیع است. با استفاده از یادگیری ماشین و پردازش زبان طبیعی، مهاجمان می‌توانند محتوای بسیار متقاعدکننده‌ای مانند اخبار جعلی، پست‌های شبکه‌های اجتماعی، جعل عمیق‌ها (تقلب‌های عمیق) یا تصاویر دستکاری‌شده ایجاد کنند که افراد، سازمان‌ها یا حتی کل جمعیت را فریب دهد. این نوع حمله به‌خصوص خطرناک است زیرا از روانشناسی انسان سوءاستفاده می‌کند، اعتماد به نهادها را تضعیف می‌کند و می‌تواند بر افکار عمومی، انتخابات یا تحولات بازار تأثیر بگذارد.

خودکارسازی و مقیاس‌پذیری تولید اطلاعات نادرست با استفاده از هوش مصنوعی، آن را به ابزاری قدرتمند برای بازیگران مخرب، از جمله گروه‌های تحت حمایت دولت، مجرمان سایبری یا سازمان‌های ایدئولوژیک تبدیل کرده است.

۲-۱۶-۱ ویژگی‌های کلیدی

- **محتوای با کیفیت بالا:**
  - هوش مصنوعی متن، صدا، ویدیو یا تصاویر واقع‌گرایانه‌ای ایجاد می‌کند که برای افراد عادی غیرقابل تشخیص از محتوای واقعی است.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند ایجاد، انتشار و تقویت اطلاعات نادرست را در چندین

پلتفرم به‌صورت خودکار انجام می‌دهند.

- **انطباق‌پذیری:**
  - سیستم‌های مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در احساسات عمومی، سیاست‌های پلتفرم‌ها یا اقدامات دفاعی سازگار شوند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی امکان تولید و توزیع سریع حجم زیادی از اطلاعات نادرست را فراهم می‌کند و تلاش‌های سنتی بررسی واقعیت‌ها را تحت فشار قرار می‌دهد.
- **دقت هدف‌گیری:**
  - هوش مصنوعی داده‌های کاربران را تحلیل می‌کند تا کمپین‌های اطلاعات نادرست را به‌طور خاص برای جمعیت‌ها، علایق یا نقاط ضعف خاص تنظیم کند.

۲-۱۶-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **پردازش زبان طبیعی:**
  - تولید متن: از مدل‌های مبتنی بر ترنسفورمر مانند GPT-3 یا BERT برای ایجاد مقالات خبری جعلی، پست‌های شبکه‌های اجتماعی یا نظرات منسجم و مرتبط با زمینه استفاده می‌کند.
  - تحلیل احساسات: احساسات عمومی را تحلیل می‌کند تا موضوعات یا روایت‌هایی را که با مخاطبان هدف طنین‌انداز می‌شوند، شناسایی کند.
  - مدل‌سازی موضوعی: موضوعات داغ یا جنجال‌ها را شناسایی می‌کند تا اطلاعات نادرست حول آن‌ها ایجاد کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های مولد تخاصمی، محتوای چندرسانه‌ای مصنوعی اما واقع‌گرایانه مانند جعل عمیق‌ها، تصاویر دستکاری‌شده یا ویدیوهای جعلی ایجاد می‌کنند.
  - شبکه‌های عصبی بازگشتی و شبکه‌های حافظه طولانی کوتاه-مدت (LSTM)[1] وابستگی‌های زمانی در مکالمات یا روایت‌ها را تحلیل می‌کنند تا استراتژی‌های اطلاعات نادرست را بهبود بخشند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار کاربران در شبکه‌های اجتماعی استفاده می‌کند و به مهاجمان امکان می‌دهد تا کاربران قانونی را تقلید کرده و از تشخیص

1. Long short-term memory (LSTM)

فرار کنند.

- **تحلیل شبکه‌های اجتماعی:**
  - شبکه‌های عصبی گرافی (GNNs)[1] ساختار شبکه‌های اجتماعی را تحلیل می‌کنند تا گره‌ها یا جوامع تأثیرگذار را برای انتشار هدف‌مند شناسایی کنند.
- **یادگیری تقویتی:**
  - کمپین‌های اطلاعات نادرست را با پاداش دادن به معیارهای موفقیت‌آمیز مشارکت (مانند لایک‌ها، اشتراک‌ها، بازدیدها) و جریمه کردن شکست‌ها بهینه می‌کند.

۳-۱۶-۲ دارایی‌های هدف

- **افکار عمومی:**
  - دستکاری باورها، ادراکات یا نگرش‌ها نسبت به شخصیت‌های سیاسی، شرکت‌ها یا مسائل اجتماعی.
- **انتخابات و حکومت:**
  - تأثیر بر نتایج انتخابات، ایجاد تفرقه میان رأی‌دهندگان یا تضعیف اعتماد به فرآیندهای دموکراتیک.
- **بازارهای مالی:**
  - انتشار اطلاعات نادرست درباره شرکت‌ها، صنایع یا شاخص‌های اقتصادی برای دستکاری قیمت سهام یا ایجاد بی‌ثباتی در بازار.
- **پلتفرم‌های رسانه‌ای:**
  - شبکه‌های اجتماعی، وب‌سایت‌های خبری یا فروم‌هایی که اطلاعات نادرست در آن‌ها از طریق الگوریتم‌های پیشنهاد یا اشتراک‌گذاری ویروسی تقویت می‌شوند.
- **زیرساخت‌های حیاتی:**
  - اختلال در اعتماد به خدمات ضروری مانند بهداشت، انرژی یا حمل‌ونقل با انتشار ترس، عدم اطمینان یا تردید.

۴-۱۶-۲ آسیب‌پذیری‌ها

- **روانشناسی انسان:**
  - افراد بیشتر احتمال دارد محتوای احساسی یا جنجالی را باور کنند، که آن‌ها را در برابر

1. Graph Neural Networks (GNNs)

اطلاعات نادرست تولیدشده توسط هوش مصنوعی آسیب‌پذیر می‌کند.

- **الگوریتم‌های اولویت‌بندی:**
  - الگوریتم‌های شبکه‌های اجتماعی محتوای جذاب را در اولویت قرار می‌دهند و به‌طور ناخواسته اطلاعات نادرستی را که واکنش‌های احساسی قوی ایجاد می‌کنند، ترویج می‌دهند.
- **مکانیزم‌های ضعیف بررسی واقعیت:**
  - منابع محدود یا تکنیک‌های قدیمی برای تأیید اعتبار محتوای آنلاین، اجازه می‌دهد اطلاعات نادرست بدون کنترل گسترش یابد.
- **نشت داده‌ها:**
  - مجموعه‌داده‌های نشت‌شده حاوی اطلاعات شخصی، مواد ارزشمندی برای ایجاد کمپین‌های اطلاعات نادرست هدف‌مند فراهم می‌کنند.
- **سوءاستفاده از پلتفرم‌ها:**
  - فقدان ابزارها یا سیاست‌های قوی مدیریت و کنترل در برخی پلتفرم‌ها، به ربات‌های خودکار و ترول‌ها اجازه می‌دهد اطلاعات نادرست را آزادانه منتشر کنند.

۵-۱۶-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره مخاطبان هدف، از جمله ترجیحات، رفتارها و آسیب‌پذیری‌های آن‌ها استفاده می‌کند.
- **تولید محتوا:**
  - مهاجم از مدل‌های پردازش زبان طبیعی یا مدل‌های زبانی بزرگ برای ایجاد اطلاعات نادرست متقاعدکننده و مرتبط با زمینه که برای گروه‌ها یا افراد خاص تنظیم شده‌اند، استفاده می‌کند.
- **ایجاد محتوای چندرسانه‌ای:**
  - مهاجم از شبکه‌های مولد تخاصمی یا سایر تکنیک‌های یادگیری عمیق و هوش مصنوعی مولد برای تولید تصاویر، ویدیوها یا کلیپ‌های صوتی مصنوعی که از روایت پشتیبانی می‌کنند، استفاده می‌شود.
- **انتشار:**
  - ارسال خودکار اطلاعات نادرست در چندین پلتفرم با استفاده از بات‌نت‌ها یا حساب‌های کاربری به‌خطرافتاده.

- **تقویت:**
  - از الگوریتم‌های شبکه‌های اجتماعی، موضوعات داغ یا شبکه‌های تأثیرگذار برای حداکثر رساندن دسترسی و تأثیر استفاده می‌شود.
- **پایداری:**
  - کمپین به‌طور مداوم بر اساس بازخوردهای لحظه‌ای اصلاح می‌شود تا تأثیر بلندمدت تضمین شود.

۲-۱۶-۶ نشانه‌های احتمال نفوذ

- **الگوهای غیرعادی محتوا:**
  - ناهنجاری‌ها در فرکانس ارسال، لحن یا سبک که از رفتار معمول کاربران منحرف می‌شود.
- **فعالیت ربات‌ها:**
  - حجم بالایی از محتوای یکسان یا مشابه که توسط چندین حساب کاربری در بازه زمانی کوتاهی ارسال می‌شود.
- **معیارهای مشارکت:**
  - افزایش ناگهانی لایک‌ها، اشتراک‌ها یا نظرات روی محتوای جنجالی یا احساسی.
- **حساب‌های کاربری غیرواقعی:**
  - پروفایل‌های مشکوک با جزئیات ناقص، تاریخ‌های ایجاد اخیر یا سطح فعالیت غیرعادی.
- **محتوای چندرسانه‌ای گمراه‌کننده:**
  - تصاویر، ویدیوها یا کلیپ‌های صوتی دستکاری‌شده یا جعلی که با واقعیت‌های تأییدشده در تضاد هستند.

۲-۱۶-۷ تأثیرات

- **تفرقه اجتماعی:**
  - قطبی‌شدن افکار عمومی، فرسایش اعتماد به نهادها و افزایش ناآرامی‌های اجتماعی.
- **پیامدهای اقتصادی:**
  - بی‌ثباتی بازار، ضررهای مالی یا آسیب به اعتبار کسب‌وکارها به‌دلیل اطلاعات نادرست.
- **تأثیرات سیاسی:**
  - دستکاری نتایج انتخابات، بی‌ثبات‌سازی دولت‌ها یا دخالت در روابط بین‌المللی.

- **خطرات بهداشت عمومی:**
  - انتشار اطلاعات نادرست درباره بحران‌های بهداشتی، واکسن‌ها یا درمان‌های پزشکی که منجر به رفتارهای مضر می‌شود.
- **چالش‌های نظارتی:**
  - مشکل در اجرای قوانین یا مقررات علیه اشکال در حال تکامل اطلاعات نادرست.

۸-۱۶-۲ راه‌های مقابله

- **تشخیص پیشرفته محتوا:**
  - از ابزارهای مبتنی بر هوش مصنوعی استفاده کنید که قادر به شناسایی محتوای مصنوعی یا دستکاری‌شده، مانند تشخیص‌دهنده‌های جعل عمیق یا ابزارهای بررسی سرقت ادبی هستند.
- **نظارت رفتاری:**
  - از هوش مصنوعی برای شناسایی الگوهای فعالیت غیرعادی که نشان‌دهنده بات‌نت‌ها یا کمپین‌های اطلاعات نادرست هماهنگ‌شده است، استفاده کنید.
- **خودکارسازی بررسی واقعیت:**
  - سیستم‌های بررسی واقعیت خودکار مبتنی بر پردازش زبان طبیعی توسعه دهید تا دقت محتوای آنلاین را در زمان واقعی تأیید کنند.
- **مدیریت پلتفرم‌ها:**
  - سیاست‌ها و ابزارهای مدیریت سخت‌گیرانه‌تری در شبکه‌های اجتماعی پیاده‌سازی کنید تا از گسترش اطلاعات نادرست جلوگیری شود.
- **آموزش کاربران:**
  - کاربران را آموزش دهید تا محتوای آنلاین را به‌طور انتقادی ارزیابی کنند، نشانه‌های اطلاعات نادرست را تشخیص دهند و فعالیت‌های مشکوک را گزارش دهند.
- **تلاش‌های مشارکتی:**
  - همکاری بین دولت‌ها، شرکت‌های فناوری و محققان را تقویت کنید تا استانداردها و راه‌حل‌های مشترک برای مقابله با اطلاعات نادرست توسعه یابد.
- **اقدامات شفافیت:**
  - از پلتفرم‌ها بخواهید در مورد تبلیغات محتوا، هدف‌گیری تبلیغات و توصیه‌های الگوریتمی شفاف‌سازی کنند.

- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی توسعه دهید و پیاده‌سازی کنید تا کمپین‌های اطلاعات نادرست را به‌سرعت شناسایی، مهار و کاهش دهید.
- **هوشمندسازی تهدیدات پیشگیرانه:**
  - از پلتفرم‌های هوشمند مبتنی بر هوش مصنوعی برای نظارت بر روندهای نوظهور اطلاعات نادرست و تکنیک‌های تطبیقی استفاده کنید.

۲-۱۷ سیبل

حمله سیبل[1] زمانی رخ می‌دهد که یک مهاجم چندین هویت یا گره جعلی در یک سیستم غیرمتمرکز، مانند شبکه بلاک‌چین، شبکه همتا به همتا (P2P) یا پلتفرم‌های اجتماعی ایجاد می‌کند. این موجودیت‌های جعلی سپس برای مختل کردن عملکرد سیستم، دستکاری مکانیزم‌های اجماع یا کسب مزایای ناعادلانه استفاده می‌شوند. در حمله سیبل مبتنی بر هوش مصنوعی، از هوش مصنوعی برای خودکارسازی ایجاد و مدیریت این هویت‌های جعلی استفاده می‌شود که این کار حمله را پیچیده‌تر، مقیاس‌پذیرتر و تشخیص آن را دشوارتر می‌کند.

هوش مصنوعی با توانایی ایجاد پروفایل‌های واقع‌گرایانه، تقلید رفتار انسانی و سازگاری پویا با اقدامات متقابل، حمله سیبل سنتی را تقویت می‌کند. این موضوع باعث می‌شود این حمله به‌ویژه در سیستم‌هایی که به تعاملات مبتنی بر اعتماد، امتیازدهی اعتبار یا تصمیم‌گیری توزیع‌شده متکی هستند، خطرناک باشد.

۲-۱۷-۱ ویژگی‌های کلیدی

- **مقیاس‌پذیری گسترده:**
  - هوش مصنوعی ایجاد تعداد زیادی هویت یا گره جعلی را خودکار می‌کند و به مهاجمان اجازه می‌دهد به سرعت سیستم را تحت فشار قرار دهند.
- **واقع‌گرایی رفتاری:**
  - هوش مصنوعی اطمینان می‌دهد که موجودیت‌های جعلی رفتار واقع‌گرایانه‌ای از خود نشان می‌دهند و تشخیص آن‌ها از کاربران یا گره‌های قانونی دشوارتر می‌شود.
- **انعطاف‌پذیری:**
  - الگوریتم‌های یادگیری ماشین به حمله اجازه می‌دهند تا با تغییرات در دفاع سیستم یا

1. Sybil Attack

الگوهای کاربران سازگار شود.

- **پنهان‌کاری:**
  - با استفاده از پردازش زبان طبیعی و مدل‌های مولد، هوش مصنوعی می‌تواند محتوا، تعاملات و پروفایل‌های متقاعدکننده‌ای ایجاد کند که به‌راحتی با فعالیت‌های واقعی ترکیب می‌شوند.
- **دستکاری هدفمند:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد روی اهداف یا مقاصد خاص، مانند تأثیرگذاری بر نتایج رأی‌گیری، کنترل تخصیص منابع یا انتشار اطلاعات نادرست، تمرکز کنند.

۲-۱۷-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را آموزش می‌دهد تا الگوهای رفتار کاربران قانونی را تشخیص دهند و هویت‌های جعلی واقع‌گرایانه ایجاد کنند.
  - یادگیری بدون نظارت: خوشه‌هایی از کاربران یا گره‌های مشابه را شناسایی می‌کند تا توزیع موجودیت‌های جعلی در شبکه را هدایت کند.
  - یادگیری تقویتی: رفتار موجودیت‌های جعلی را بر اساس بازخورد محیط بهینه می‌کند و اثربخشی آن‌ها را در طول زمان بهبود می‌بخشد.
- **پردازش زبان طبیعی:**
  - متن واقع‌گرایانه برای پست‌های شبکه‌های اجتماعی، نظرات یا پیام‌های چت ایجاد می‌کند تا حساب‌های جعلی شبیه‌تر به انسان به نظر برسند.
  - الگوهای زبانی را تحلیل می‌کند تا محتوا را برای مخاطبان یا زمینه‌های خاص تنظیم کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های مولد تخاصمی، پروفایل‌ها، تصاویر یا الگوهای رفتاری مصنوعی اما باورپذیر ایجاد می‌کنند.
  - شبکه‌های عصبی بازگشتی داده‌های ترتیبی، مانند تاریخچه تعاملات، را مدل‌سازی می‌کنند تا فعالیت کاربران واقعی را شبیه‌سازی کنند.
- **بینایی کامپیوتر:**
  - برای ایجاد یا دستکاری تصاویر پروفایل، آواتارها یا سایر عناصر بصری مرتبط با هویت‌های جعلی استفاده می‌شود.

- **تحلیل شبکه‌های اجتماعی:**
  - از نظریه گراف و تکنیک‌های هوش مصنوعی برای تحلیل روابط بین کاربران یا گره‌ها استفاده می‌کند و امکان قرارگیری استراتژیک موجودیت‌های جعلی در شبکه را فراهم می‌کند.

۲-۱۷-۳ دارایی‌های هدف

- **شبکه‌های بلاک‌چین:**
  - دفترکل‌های غیرمتمرکز که به مکانیزم‌های اجماع مانند Proof-of-Work (PoW) یا Proof-of-Stake (PoS) متکی هستند، می‌توانند با معرفی گره‌های جعلی دستکاری شوند.
- **شبکه‌های همتا به همتا (P2P):**
  - پلتفرم‌های اشتراک‌گذاری فایل، سیستم‌های ارتباطی یا برنامه‌های مشارکتی می‌توانند با معرفی همتایان مخرب که منابع را مصرف می‌کنند یا داده‌های خراب را منتشر می‌کنند، مختل شوند.
- **پلتفرم‌های شبکه‌های اجتماعی:**
  - حساب‌های جعلی می‌توانند برای تقویت تبلیغات، انتشار اطلاعات نادرست یا دستکاری افکار عمومی در انتخابات یا بحران‌ها استفاده شوند.
- **سیستم‌های جمع‌سپاری:**
  - پلتفرم‌هایی مانند Amazon Mechanical Turk یا Reddit می‌توانند با ایجاد مشارکت‌کنندگان جعلی برای تحریف نتایج یا سوءاستفاده از سیستم‌های پاداش، مورد سوءاستفاده قرار گیرند.
- **شبکه‌های اینترنت اشیا (IoT):**
  - اکوسیستم‌های توزیع‌شده اینترنت اشیا می‌توانند با معرفی دستگاه‌های مخرب که در جمع‌آوری داده‌ها، پردازش یا فعال‌سازی اختلال ایجاد می‌کنند، به خطر بیفتند.
- **سیستم‌های اعتبارسنجی:**
  - بازارهای آنلاین، پلتفرم‌های بررسی یا موتورهای توصیه‌گر می‌توانند با هجوم رتبه‌بندی‌ها یا نظرات جعلی دستکاری شوند.

۲-۱۷-۴ آسیب‌پذیری‌ها

- **تأیید هویت ضعیف:**
  - سیستم‌هایی که به فرآیندهای ثبت‌نام ساده متکی هستند یا فاقد بررسی‌های هویتی قوی

هستند، آسیب‌پذیرترند.

- **تشخیص ناهنجاری ناکافی:**
  - عدم وجود ابزارهای نظارتی پیشرفته برای شناسایی الگوهای فعالیت غیرعادی یا خوشه‌بندی حساب‌های مشابه.
- **شکاف‌های مقیاس‌پذیری:**
  - شبکه‌هایی که قادر به مدیریت ورود تعداد زیادی شرکت‌کننده جدید نیستند، ممکن است در تشخیص تفاوت بین موجودیت‌های قانونی و جعلی مشکل داشته باشند.
- **عدم تحلیل رفتاری:**
  - سیستم‌هایی که رفتار کاربران را در طول زمان تحلیل نمی‌کنند، ممکن است نتوانند ناسازگاری‌ها یا انحرافات نشان‌دهنده حملات سیبل را تشخیص دهند.
- **خطای انسانی:**
  - تنظیمات نادرست کنترل‌های دسترسی یا نظارت ناکافی ممکن است به‌طور ناخواسته معرفی هویت‌های جعلی را تسهیل کند.

۵-۱۷-۲ سناریوی تهدید

- **ایجاد هویت:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای ایجاد تعداد زیادی هویت جعلی، همراه با پروفایل‌ها، تصاویر و الگوهای رفتاری واقع‌گرایانه استفاده می‌کند.
- **نفوذ به شبکه:**
  - مهاجم موجودیت‌های جعلی را در سیستم هدف مستقر می‌کند و اطمینان حاصل می‌کند که به‌طور استراتژیک توزیع شده‌اند تا حداکثر تأثیر را داشته باشند.
- **شبیه‌سازی رفتار:**
  - مهاجم از مدل‌های یادگیری ماشین برای شبیه‌سازی رفتار کاربران واقعی، مانند ارسال محتوا، تعامل با دیگران یا مشارکت در مکانیزم‌های اجماع استفاده می‌کند.
- **مصرف منابع:**
  - مهاجم از وجود موجودیت‌های جعلی برای مصرف بیش از حد منابع، اختلال در دسترسی به خدمات یا دستکاری فرآیندهای تصمیم‌گیری سوءاستفاده می‌کند.

- **دستکاری و کنترل:**
  - مهاجم از هویت‌های جعلی برای تأثیرگذاری بر نتایج، مانند تغییر نتایج رأی‌گیری، انتشار اطلاعات نادرست یا کسب کنترل نامتناسب بر منابع مشترک استفاده می‌کند.

۲-۱۷-۶ نشانه‌های احتمال نفوذ

- **رشد غیرعادی حساب‌ها:**
  - افزایش ناگهانی تعداد حساب‌ها یا گره‌های جدید که در بازه زمانی کوتاهی به سیستم می‌پیوندند.
- **رفتار خوشه‌ای:**
  - گروه‌هایی از حساب‌ها یا گره‌ها که رفتار بسیار همبسته‌ای از خود نشان می‌دهند، مانند الگوهای ارسال یکسان یا اقدامات هماهنگ.
- **الگوهای فعالیت غیرعادی:**
  - حساب‌هایی که سطح فعالیت غیرمعمول بالایی دارند یا اقداماتی انجام می‌دهند که با رفتار معمول کاربران ناسازگار است.
- **ویژگی‌های مشترک:**
  - چندین حساب که ویژگی‌های مشترکی مانند آدرس‌های IP، دامنه‌های ایمیل یا جزئیات پروفایل دارند.
- **محتواهای غیرواقعی:**
  - پست‌ها، نظرات یا پیام‌های تولیدشده توسط حساب‌های جعلی که فاقد عمق، زمینه یا انسجام هستند.
- **اضافه‌بار منابع:**
  - نشانه‌هایی از افزایش بار روی سیستم، مانند زمان‌های پاسخ کندتر یا استفاده بیشتر از پهنای باند، بدون تقاضای قانونی متناظر.

۲-۱۷-۷ تأثیرات

- **اختلال در اجماع:**
  - دستکاری فرآیندهای رأی‌گیری یا تصمیم‌گیری می‌تواند منجر به نتایج نادرست یا فلج شدن سیستم شود.

- **اتمام منابع:**
  - مصرف بیش از حد منابع محاسباتی، ذخیره‌سازی یا شبکه‌ای می‌تواند عملکرد را کاهش دهد و هزینه‌های عملیاتی را افزایش دهد.
- **از دست دادن اعتماد:**
  - کاربران ممکن است در صورت مشاهده دستکاری گسترده یا بی‌عدالتی، اعتماد خود را به سیستم از دست بدهند.
- **ضررهای مالی:**
  - فعالیت‌های کلاهبرداری، مانند double-spending در شبکه‌های بلاک‌چین یا سوءاستفاده از سیستم‌های پاداش، می‌تواند منجر به ضررهای مالی قابل توجهی شود.
- **آسیب به اعتبار:**
  - سازمان‌های استفاده‌کننده از پلتفرم‌های آسیب‌دیده ممکن است به دلیل سهل‌انگاری یا آسیب‌پذیری، آسیب اعتباری ببینند.
- **خطرات امنیتی:**
  - حملات سیبل موفق می‌توانند به عنوان نقطه ورود برای نفوذهای شدیدتر، مانند نقض داده‌ها یا انتشار بدافزار، عمل کنند.

۸-۱۷-۲ راه‌های مقابله

- **تأیید هویت قوی:**
  - احراز هویت چندعاملی، چالش‌های CAPTCHA یا تأیید بیومتریک را برای اطمینان از قانونی بودن حساب‌های جدید پیاده‌سازی کنید.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای ایجاد خطوط پایه و تشخیص ناهنجاری‌های نشان‌دهنده فعالیت سیبل استفاده کنید.
- **سیستم‌های اعتبارسنجی:**
  - مکانیزم‌های مبتنی بر اعتبار را معرفی کنید که در آن حساب‌های قدیمی‌تر و معتبرتر تأثیر بیشتری نسبت به حساب‌های جدیدتر دارند.
- **محدود کردن نرخ:**
  - تعداد اقداماتی که یک حساب می‌تواند در یک بازه زمانی مشخص انجام دهد را محدود کنید تا از سوءاستفاده جلوگیری شود.

- **مدل‌های اعتماد غیرمتمرکز:**
  - از چارچوب‌های اعتمادی استفاده کنید که برای تأیید موجودیت‌ها یا تراکنش‌های جدید به همکاری چندین طرف نیاز دارند.
- **بررسی‌های منظم:**
  - ممیزی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **گزارش‌دهی جامعه:**
  - کاربران را تشویق کنید تا فعالیت‌های مشکوک را گزارش دهند و برای شناسایی حملات سیبل بالقوه مشوق ارائه دهید.
- **استفاده از الگوریتم‌های پیشرفته:**
  - الگوریتم‌هایی را پیاده‌سازی کنید که به‌طور خاص برای تشخیص و کاهش حملات سیبل طراحی شده‌اند، مانند SybilLimit ، SybilGuard یا EigenTrust.

۲-۱۸ فیشینگ هدفمند

حمله فیشینگ هدفمند[1] نوعی حمله فیشینگ است که در آن مهاجم پیام‌های بسیار شخصی‌سازی‌شده و متقاعدکننده‌ای را به افراد یا سازمان‌های خاص ارسال می‌کند تا اطلاعات حساس، اعتبارنامه‌ها یا داده‌های مالی را سرقت کند. در حمله فیشینگ هدفمند مبتنی بر هوش مصنوعی، از هوش مصنوعی برای افزایش دقت، شخصی‌سازی و مقیاس‌پذیری این حملات استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و پردازش زبان طبیعی، مهاجمان می‌توانند حجم عظیمی از داده‌ها را از شبکه‌های اجتماعی، تاریخچه ایمیل‌ها و سوابق عمومی تحلیل کنند تا پیام‌های فیشینگ بسیار متقاعدکننده و آگاه از زمینه ایجاد کنند.

این نوع حمله از روان‌شناسی انسان سوءاستفاده می‌کند و احتمال فریب خوردن قربانیان را افزایش می‌دهد. حملات فیشینگ هدفمند مبتنی بر هوش مصنوعی می‌توانند از فیلترهای ایمیل و سیستم‌های تشخیص اسپام عبور کنند و نرخ موفقیت خود را افزایش دهند.

۲-۱۸-۱ ویژگی‌های کلیدی

- **محتوای بسیار شخصی‌سازی‌شده:**
  - هوش مصنوعی حضور آنلاین، الگوهای ارتباطی و علایق هدف را تحلیل می‌کند تا پیام‌های

1. Spear Phishing

فیشینگ سفارشی‌سازی‌شده ایجاد کند که قانونی به نظر می‌رسند.

- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند شناسایی اهداف، ایجاد پیام‌ها و اجرای حملات را خودکار می‌کنند و تلاش دستی را کاهش می‌دهند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در رفتار کاربر، اقدامات امنیتی یا فناوری‌های دفاعی سازگار شوند.
- **پنهان‌کاری:**
  - با تقلید از سبک‌های ارتباطی قانونی و اجتناب از کلمات کلیدی مشکوک، حملات فیشینگ هدفمند مبتنی بر هوش مصنوعی می‌توانند از تشخیص توسط فیلترهای ایمیل و سیستم‌های امنیتی اجتناب کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین فرد یا سازمان را به‌طور همزمان هدف‌گیری کنند، که این امر دامنه و تأثیر حمله را افزایش می‌دهد.

۲-۱۸-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌های داده‌ای از ایمیل‌های فیشینگ موفق آموزش می‌دهد تا الگوها را یاد بگیرد و تولید پیام را بهبود بخشد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در داده‌های ارتباطی شناسایی می‌کند تا اهداف بالقوه را کشف کند یا استراتژی‌های حمله را بهبود بخشد.
  - یادگیری تقویتی: فرآیند فیشینگ هدفمند را با پاداش دادن به تعامل‌های موفقیت‌آمیز و جریمه کردن شکست‌ها بهینه می‌کند.
- **پردازش زبان طبیعی:**
  - الگوهای زبانی، لحن و سبک را در ارتباطات قانونی تحلیل می‌کند تا پیام‌های فیشینگ واقع‌گرایانه و آگاه از زمینه ایجاد کند.
  - اطلاعات مرتبط را از شبکه‌های اجتماعی، آرشیو ایمیل‌ها و سایر منابع استخراج می‌کند تا محتوا را شخصی‌سازی کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، متن‌های منسجم و متقاعدکننده ایجاد می‌کنند که سبک نوشتاری مخاطبان یا مؤسسات مورد اعتماد را تقلید می‌کنند.
  - شبکه‌های مولد تخاصمی: قالب‌های ایمیل، پیوست‌ها یا حتی شخصیت‌های جعلی مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قربانیان را فریب دهند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر استفاده می‌کند و به مهاجمان امکان می‌دهد تا زمان و نحوه تعامل با اهداف را برای حداکثر تأثیر پیش‌بینی کنند.
- **تحلیل شبکه‌های اجتماعی:**
  - از هوش مصنوعی برای جمع‌آوری و تحلیل داده‌های عمومی از پلتفرم‌هایی مانند LinkedIn Twitter یا Facebook استفاده می‌کند تا بینش‌هایی درباره زندگی حرفه‌ای و شخصی هدف به‌دست آورد.

۳-۱۸-۲ دارایی‌های هدف

- **افراد:**
  - مدیران، کارمندان یا پرسنل کلیدی که دسترسی آن‌ها به سیستم‌ها یا داده‌های حساس، آن‌ها را به اهداف ارزشمندی تبدیل می‌کند.
- **سازمان‌ها:**
  - شرکت‌ها، سازمان‌های دولتی یا مؤسساتی که مالکیت فکری، اسرار تجاری یا داده‌های مشتری را ذخیره می‌کنند.
- **سیستم‌های مالی:**
  - بانک‌ها، شرکت‌های کارت اعتباری و پلتفرم‌های پرداخت آنلاین که اعتبارنامه‌های سرقت‌شده می‌توانند برای کسب سود مالی سوءاستفاده شوند.
- **زنجیره‌های تأمین:**
  - فروشندگان، شرکا یا تأمین‌کنندگان شخص ثالث که حساب‌های به‌خطر افتاده آن‌ها می‌توانند به عنوان نقاط ورود به اکوسیستم‌های بزرگتر عمل کنند.

- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تأسیسات بهداشتی یا شبکه‌های حمل‌ونقل که فیشینگ هدفمند می‌تواند منجر به اختلالات قابل توجهی شود.

۲-۱۸-۴ آسیب‌پذیری‌ها

- **روان‌شناسی انسان:**
  - افراد بیشتر احتمال دارد به پیام‌های شخصی‌سازی‌شده و آگاه از زمینه اعتماد کنند، که آن‌ها را در برابر تلاش‌های فیشینگ مبتنی بر هوش مصنوعی آسیب‌پذیر می‌کند.
- **نشت داده:**
  - پایگاه‌های داده نشت‌یافته حاوی اطلاعات شخصی، مواد ارزشمندی برای حملات فیشینگ هدفمند مبتنی بر هوش مصنوعی فراهم می‌کنند.
- **امنیت ضعیف ایمیل:**
  - عدم وجود مکانیزم‌های فیلترسازی، رمزنگاری یا احراز هویت قوی ایمیل، کاربران را در معرض پیام‌های مخرب قرار می‌دهد.
- **آموزش ناکافی:**
  - کارمندان یا کاربرانی که برای تشخیص تلاش‌های فیشینگ پیچیده آموزش ندیده‌اند، ممکن است قربانی این حملات شوند.
- **اتکای بیش‌ازحد به خودکارسازی:**
  - فیلترهای ایمیل و سیستم‌های تشخیص هرزنامه سنتی ممکن است نتوانند محتوای تولیدشده توسط هوش مصنوعی را به دلیل سطح بالای شخصی‌سازی و واقع‌گرایی تشخیص دهند.

۲-۱۸-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره هدف، از جمله نقش شغلی، روابط، عادات ارتباطی و فعالیت‌های اخیر استفاده می‌کند.
- **جمع‌آوری داده:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های عمومی، مانند پست‌های شبکه‌های اجتماعی، آرشیو ایمیل‌ها یا مقالات خبری مستقر می‌کند تا پروفایل دقیقی از هدف ایجاد کند.

- **تولید پیام:**
  - مهاجم از تکنیک‌های پردازش زبان طبیعی و یادگیری عمیق برای ایجاد پیام‌های فیشینگ شخصی‌سازی‌شده و متقاعدکننده استفاده می‌کند که سبک ارتباطات قانونی را تقلید می‌کنند.
- **استقرار:**
  - مهاجم پیام‌های ایجادشده را به هدف ارسال می‌کند و اطمینان می‌یابد که به نظر می‌رسد از یک منبع مورد اعتماد، مانند همکار، فروشنده یا مؤسسه ارسال شده‌اند.
- **سوءاستفاده:**
  - مهاجم هنگامی که قربانی با پیام فیشینگ تعامل می‌کند (مانند کلیک روی لینک، دانلود پیوست یا ارائه اعتبارنامه)، از نفوذ برای کسب سود مالی، سرقت داده یا نفوذ بیشتر استفاده می‌کند.
- **پایداری:**
  - مهاجم دسترسی به سیستم‌ها یا حساب‌های به‌خطر افتاده را با بهبود مداوم استراتژی حمله بر اساس نتایج مشاهده‌شده حفظ می‌کند.

۶-۱۸-۲ نشانه‌های احتمال نفوذ

- **رفتار غیرعادی فرستنده:**
  - ایمیل‌هایی از دامنه‌هایی که به نظر قانونی می‌رسند اما کمی تغییر یافته‌اند یا از فرستنده‌های غیرمنتظره ارسال شده‌اند.
- **لینک‌ها یا پیوست‌های مشکوک:**
  - پیام‌های حاوی URLهای کوتاه‌شده، انواع فایل ناشناخته یا درخواست‌هایی برای اطلاعات حساس.
- **درخواست‌های غیرمنتظره:**
  - درخواست‌های فوری یا خارج از عرف برای پول، اعتبارنامه یا داده‌های محرمانه.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر ناسازگار است، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به احراز هویت ناموفق، قفل شدن

حساب‌ها یا فعالیت‌های غیرمنتظره.

۲-۱۸-۷ تأثیرات

- **نشت داده:**
  - سرقت اطلاعات حساس، مانند مالکیت فکری، اسرار تجاری یا داده‌های مشتری، که منجر به خسارات مالی یا آسیب به اعتبار می‌شود.
- **خسارات مالی:**
  - تراکنش‌های غیرمجاز، انتقال‌های بانکی کلاهبرداری یا عفونت‌های باج‌افزاری که باعث آسیب مالی می‌شوند.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل به‌خطر افتادن اعتبارنامه‌ها یا سوءاستفاده از آسیب‌پذیری‌ها.
- **آسیب به اعتبار:**
  - از دست دادن اعتماد مشتری و ارزش برند پس از نفوذ یا سوءاستفاده از هویت‌های سرقت‌شده.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

۲-۱۸-۸ راه‌های مقابله

- **فیلترسازی ایمیل:**
  - راه‌حل‌های پیشرفته فیلترسازی ایمیل را پیاده‌سازی کنید که قادر به تشخیص تلاش‌های فیشینگ مبتنی بر هوش مصنوعی با استفاده از تحلیل‌های رفتاری و تحلیل زمینه‌ای هستند.
- **احراز هویت چندعاملی:**
  - مراحل تأیید اضافی مانند کدهای یک‌بارمصرف یا اسکن‌های بیومتریک را فراتر از رمز عبور الزامی کنید تا از دسترسی غیرمجاز جلوگیری شود.
- **آموزش آگاهی امنیتی:**
  - کارمندان و کاربران را در مورد تشخیص علائم فیشینگ هدفمند، مانند لینک‌های مشکوک، درخواست‌های غیرمنتظره یا جزئیات تغییر یافته فرستنده آموزش دهید.

- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت که نشان‌دهنده تلاش‌های فیشینگ هدفمند یا حساب‌های به‌خطر افتاده هستند، استفاده کنید.
- **نظارت بر دامنه:**
  - برای جعل دامنه یا تلاش‌های typosquatting که دامنه‌های قانونی را تقلید می‌کنند، نظارت کرده و آن‌ها را مسدود کنید.
- **رمزنگاری:**
  - ارتباطات و داده‌های حساس را رمزنگاری کنید تا از رهگیری و سوءاستفاده جلوگیری شود.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **برنامه پاسخ به حوادث:**
  - یک برنامه پاسخ به حوادث قوی ایجاد و پیاده‌سازی کنید تا به‌سرعت حملات فیشینگ هدفمند موفق را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.
- **معماری Zero-Trust:**
  - یک رویکرد zero-trust را برای امنیت شبکه اتخاذ کنید که هر تلاش دسترسی را بدون توجه به منشأ آن تأیید می‌کند.

۲-۱۹ حدس رمز عبور

حمله حدس رمز عبور[1] شامل تلاش برای کشف رمز عبور کاربر با استفاده از حدس‌های آگاهانه بر اساس الگوهای رایج، اطلاعات شخصی یا داده‌های نشت‌یافته است. در حالی که حدس رمز عبور سنتی به لیست‌های از پیش تعریف‌شده یا تلاش‌های دستی متکی است، حمله حدس رمز عبور مبتنی بر هوش مصنوعی از هوش مصنوعی برای تحلیل مجموعه‌های بزرگ داده، شناسایی روندها و تولید حدس‌های بسیار هدفمند استفاده می‌کند. این امر باعث می‌شود حمله کارآمدتر، انعطاف‌پذیرتر و قادر به دور زدن دفاع‌های ساده مانند محدودیت نرخ یا قفل شدن حساب‌ها باشد.

با استفاده از یادگیری ماشین و پردازش زبان طبیعی، مهاجمان می‌توانند مدل‌های پیچیده‌ای ایجاد کنند که رمزهای عبور را با دقت بیشتری پیش‌بینی کنند، به ویژه زمانی که با اطلاعات زمینه‌ای مانند

1. Password guessing attack

جمعیت‌شناسی کاربر، داده‌های رفتاری یا اعتبارنامه‌های نشت‌یافته قبلی ترکیب شوند.

۲-۱۹-۱ ویژگی‌های کلیدی

- **پیش‌بینی مبتنی بر داده:**
  - هوش مصنوعی مجموعه‌های بزرگ داده‌های رمزهای عبور نشت‌یافته، الگوهای زبانی و رفتار کاربر را تحلیل می‌کند تا رمزهای عبور محتمل را پیش‌بینی کند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند تولید و آزمایش حدس‌های رمز عبور را خودکار می‌کنند و تلاش دستی را کاهش می‌دهند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند به صورت پویا با تغییرات در سیاست‌های رمز عبور، الگوریتم‌های هش یا مکانیزم‌های دفاعی مانند CAPTCHA یا احراز هویت چندعاملی سازگار شوند.
- **آگاهی زمینه‌ای:**
  - با گنجاندن اطلاعات زمینه‌ای (مانند نام کاربری، مکان، علایق)، هوش مصنوعی می‌تواند حدس‌های خود را برای هدف‌گیری کاربران خاص بهینه کند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین حساب یا سیستم را به طور همزمان هدف‌گیری کنند، که این امر دامنه و تأثیر حمله را افزایش می‌دهد.

۲-۱۹-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌های داده برچسب‌خورده از رمزهای عبور نشت‌یافته آموزش می‌دهد تا الگوها و ساختارهای رایج را یاد بگیرد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در مجموعه‌های داده رمز عبور بدون برچسب‌گذاری قبلی شناسایی می‌کند.
  - یادگیری تقویتی: با پاداش دادن به حدس‌های موفق و جریمه کردن شکست‌ها، استراتژی‌های حمله را بهینه می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی بازگشتی و شبکه‌های حافظه طولانی کوتاه-مدت (LSTM)، وابستگی‌های ترتیبی در رمزهای عبور را تحلیل می‌کنند تا ترکیبات محتمل را پیش‌بینی کنند.
  - شبکه‌های مولد تخاصمی: رمزهای عبور مصنوعی اما واقع‌گرایانه بر اساس الگوهای یادگرفته‌شده از داده‌های دنیای واقعی ایجاد می‌کنند.
- **پردازش زبان طبیعی:**
  - الگوهای زبانی در رمزهای عبور، مانند ساختار کلمات، استفاده از حروف بزرگ و کاراکترهای خاص را تحلیل می‌کند تا حدس‌های دقیق‌تری ایجاد کند.
  - از پردازش زبان طبیعی برای گنجاندن سرنخ‌های زمینه‌ای مانند نام کاربری، مکان یا علایق شخصی در پیش‌بینی رمز عبور استفاده می‌کند.
- **تحلیل آماری:**
  - از نظریه احتمال و استنتاج آماری برای تخمین احتمال اجزای خاص رمز عبور بر اساس داده‌های مشاهده‌شده استفاده می‌کند.
- **الگوریتم‌های تکاملی:**
  - از تکنیک‌های تکاملی برای بهبود تکراری حدس‌های رمز عبور با شبیه‌سازی فرآیندهای انتخاب طبیعی استفاده می‌کند.

۳-۱۹-۲ دارایی‌های هدف

- **حساب‌های کاربری:**
  - حساب‌های ایمیل، پروفایل‌های شبکه‌های اجتماعی، پلتفرم‌های بانکداری آنلاین و برنامه‌های سازمانی که اطلاعات حساس را ذخیره می‌کنند.
- **خدمات ابری:**
  - برنامه‌ها و APIهای میزبانی‌شده در ابر که در معرض اینترنت قرار دارند و رمزهای عبور ضعیف می‌توانند منجر به دسترسی غیرمجاز شوند.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - لوازم خانگی هوشمند، دستگاه‌های پوشیدنی و سنسورهای صنعتی که اغلب با رمزهای عبور پیش‌فرض یا ضعیف پیکربندی شده‌اند.

- **سیستم‌های سازمانی:**
  - شبکه‌های شرکتی، سرورهای فایل و سیستم‌های پایگاه‌داده که داده‌های حیاتی کسب‌وکار را ذخیره می‌کنند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، سیستم‌های تأمین آب و شبکه‌های حمل‌ونقل که برای عملکرد به احراز هویت امن متکی هستند.

۲-۱۹-۴ آسیب‌پذیری‌ها

- **رمزهای عبور قابل پیش‌بینی:**
  - انتخاب رمزهای عبور رایج یا به راحتی قابل حدس توسط کاربران، آن‌ها را در برابر حملات مبتنی بر هوش مصنوعی آسیب‌پذیر می‌کند.
- **سیاست‌های ضعیف رمز عبور:**
  - عدم وجود الزامات پیچیدگی یا استفاده مجدد از رمزهای عبور در چندین حساب، خطر نفوذ را افزایش می‌دهد.
- **اعتبارنامه‌های نشت‌یافته:**
  - نشت داده‌ها، اطلاعات ارزشمندی را برای آموزش حملات مبتنی بر هوش مصنوعی فراهم می‌کند و به مهاجمان امکان می‌دهد استراتژی‌های خود را بهبود بخشند.
- **محدودیت نرخ ناکافی:**
  - عدم وجود مکانیزم‌های محدودیت نرخ مؤثر، به مهاجمان امکان می‌دهد تا تلاش‌های ورود سریع را بدون ایجاد هشدار انجام دهند.
- **خطای انسانی:**
  - کاربرانی که اطلاعات شخصی خود را به صورت عمومی به اشتراک می‌گذارند (مانند شبکه‌های اجتماعی) یا احراز هویت چندعاملی را فعال نمی‌کنند، خطر نفوذ را افزایش می‌دهند.

۲-۱۹-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم با استفاده از ابزارهای مبتنی بر هوش مصنوعی، مجموعه‌های بزرگ داده‌های رمزهای عبور نشت‌یافته، رفتار کاربر و اطلاعات زمینه‌ای را جمع‌آوری می‌کند.

- **آموزش مدل:**
  - مهاجم مدل‌های یادگیری ماشین را بر روی داده‌های جمع‌آوری‌شده آموزش می‌دهد تا ساختارها، الگوها و نقاط ضعف رایج رمز عبور را شناسایی کند.
- **تحلیل زمینه‌ای:**
  - مهاجم اطلاعات زمینه‌ای اضافی مانند نام کاربری، مکان یا علایق شخصی را برای بهبود حدس‌های رمز عبور در نظر می‌گیرد.
- **اجرای حمله:**
  - مدل آموزش‌دیده را برای تولید و آزمایش حدس‌های رمز عبور در برابر سیستم هدف مستقر می‌کند و کاندیداهای با احتمال بالا را اولویت‌بندی می‌کند.
- **سازگاری:**
  - مهاجم استراتژی حمله را بر اساس بازخورد تلاش‌های قبلی به‌طور مداوم بهبود می‌بخشد و با تغییرات در سیاست‌های رمز عبور یا اقدامات دفاعی سازگار می‌شود.
- **سوءاستفاده:**
  - مهاجم پس از یافتن یک رمز عبور معتبر، از آن برای دسترسی غیرمجاز به حساب یا سیستم هدف استفاده می‌کند.

۶-۱۹-۲ نشانه‌های احتمال نفوذ

- **تلاش‌های ناموفق ورود:**
  - افزایش ناگهانی در تلاش‌های ناموفق ورود به دنبال دسترسی موفق ممکن است نشان‌دهنده حمله بروت فورس باشد.
- **الگوهای فعالیت غیرعادی:**
  - ناهنجاری‌ها در زمان‌های ورود، مکان‌ها یا دستگاه‌هایی که با کاربران قانونی مرتبط نیستند.
- **استفاده افزایش‌یافته از منابع:**
  - مصرف بیشتر CPU یا حافظه به دلیل تلاش‌های مکرر ورود یا سایر فعالیت‌های خودکار.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر ناسازگار است، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.

- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مرتبط با شکست‌های احراز هویت، قفل شدن حساب‌ها یا فعالیت غیرمنتظره.

۲-۱۹-۷ تأثیرات

- **دسترسی غیرمجاز:**
  - حدس موفقیت‌آمیز منجر به دسترسی غیرمجاز به حساب‌ها یا سیستم‌های حساس می‌شود و امکان سوءاستفاده بیشتر را فراهم می‌کند.
- **نقض داده‌ها:**
  - سرقت داده‌های شخصی، مالی یا مالکیت فکری که منجر به ضررهای مالی قابل توجه یا آسیب به اعتبار می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل نفوذ به اعتبارنامه‌ها.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلال در خدمات منجر به ضررهای مالی می‌شود.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

۲-۱۹-۸ راه‌های مقابله

- **سیاست‌های قوی رمز عبور:**
  - الزامات پیچیدگی را اعمال کنید، رمزهای عبور رایج را ممنوع کنید و به‌روزرسانی‌های منظم را اجباری کنید تا خطر حدس زدن کاهش یابد.
- **احراز هویت چندعاملی:**
  - مراحل تأیید اضافی مانند کدهای یک‌بارمصرف یا اسکن‌های بیومتریک را فراتر از رمز عبور الزامی کنید تا از دسترسی غیرمجاز جلوگیری شود.
- **محدودیت نرخ:**
  - مکانیزم‌های محدودیت نرخ سختگیرانه را پیاده‌سازی کنید تا تعداد تلاش‌های ورود در واحد

زمان را محدود کرده و برای فعالیت‌های مشکوک هشدار ایجاد کنید.

- **قفل شدن حساب‌ها:**
  - پس از تعداد مشخصی تلاش ناموفق ورود، حساب‌ها را به‌طور موقت غیرفعال کنید تا از حدس زدن مداوم جلوگیری شود.
- **هش کردن رمز عبور:**
  - از الگوریتم‌های هش قوی و نمک‌دار مانند Argon2، bcrypt یا PBKDF2 برای ذخیره ایمن رمزهای عبور و کند کردن تلاش‌های شکستن آن‌ها استفاده کنید.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای ورود غیرمعمول یا فعالیت‌های مشکوک که نشان‌دهنده حملات حدس رمز عبور هستند، استفاده کنید.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد بهترین روش‌ها برای ایجاد رمزهای عبور قوی و تشخیص علائم نفوذ آموزش دهید.
- **CAPTCHA و تشخیص ربات:**
  - چالش‌های CAPTCHA یا سیستم‌های پیشرفته تشخیص ربات را پیاده‌سازی کنید تا از تلاش‌های خودکار حدس زدن جلوگیری شود.

۲-۲۰ بروت فورس رمز عبور

حمله بروت فورس رمز عبور[1] شامل تلاش سیستماتیک برای امتحان کردن تمام ترکیبات ممکن از کاراکترها تا زمانی که رمز عبور صحیح پیدا شود، است. حملات بروت فورس سنتی از نظر محاسباتی بسیار سنگین و زمان‌بر هستند، به‌ویژه زمانی که با رمزهای عبور پیچیده یا الگوریتم‌های هش قوی سروکار داریم. با این حال، در حمله بروت فورس رمز عبور مبتنی بر هوش مصنوعی، از هوش مصنوعی برای بهینه‌سازی فرآیند با پیش‌بینی الگوهای احتمالی رمز عبور، کاهش فضای جستجو و سازگاری با اقدامات امنیتی مدرن مانند محدودیت نرخ یا قفل شدن حساب‌ها استفاده می‌شود.

1. Password brute-force attack

با استفاده از یادگیری ماشین و پردازش زبان طبیعی، هوش مصنوعی می‌تواند مجموعه‌داده‌های بزرگی از رمزهای عبور لورفته، رفتار کاربران و الگوهای زبانی را تحلیل کند تا حدس‌های بسیار هدف‌مندی ایجاد کند که احتمال موفقیت را به‌طور قابل توجهی افزایش می‌دهد.

۲-۲۰-۱ ویژگی‌های کلیدی

- **حدس‌زنی هوشمند:**
  - مدل‌های هوش مصنوعی پایگاه‌داده‌های رمزهای عبور لورفته، لیست‌های کلمات رایج و رفتار کاربران را تحلیل می‌کنند تا رمزهای عبور احتمالی را به‌طور کارآمدتر از حدس‌های تصادفی پیش‌بینی کنند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند تولید و آزمایش ترکیبات رمز عبور را خودکار می‌کنند و تلاش دستی را کاهش می‌دهند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند به‌طور پویا با تغییرات در سیاست‌های رمز عبور، الگوریتم‌های هش یا مکانیزم‌های دفاعی مانند محدودیت نرخ سازگار شوند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین حساب یا سیستم را به‌طور همزمان هدف قرار دهند و دامنه و تأثیر حمله را افزایش دهند.
- **پنهان‌کاری:**
  - با تقلید از تلاش‌های ورود شبیه به انسان و اجتناب از الگوهای مشکوک، حملات مبتنی بر هوش مصنوعی می‌توانند از تشخیص توسط سیستم‌های تشخیص نفوذ (IDS) فرار کنند.

۲-۲۰-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های برچسب‌دار از رمزهای عبور لورفته آموزش می‌دهد تا الگوها و ساختارهای رایج را یاد بگیرد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در مجموعه‌داده‌های رمز عبور بدون برچسب‌گذاری قبلی شناسایی می‌کند.
  - یادگیری تقویتی: استراتژی‌های حمله را با پاداش دادن به حدس‌های موفق و جریمه کردن

شکست‌ها بهینه می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و شبکه‌های حافظه کوتاه‌مدت بلندمدت، وابستگی‌های ترتیبی در رمزهای عبور را تحلیل می‌کنند تا ترکیبات احتمالی را پیش‌بینی کنند.
  - شبکه‌های مولد تخاصمی: رمزهای عبور مصنوعی اما واقع‌گرایانه بر اساس الگوهای یادگرفته‌شده از داده‌های دنیای واقعی ایجاد می‌کنند.
- **پردازش زبان طبیعی:**
  - الگوهای زبانی در رمزهای عبور، مانند ساختار کلمات، استفاده از حروف بزرگ و کاراکترهای خاص، را تحلیل می‌کند تا حدس‌های دقیق‌تری ایجاد کند.
- **تحلیل آماری:**
  - از نظریه احتمال و استنتاج آماری برای تخمین احتمال اجزای خاص رمز عبور بر اساس داده‌های مشاهده‌شده استفاده می‌کند.
- **الگوریتم‌های ژنتیکی:**
  - از تکنیک‌های تکاملی برای بهبود تدریجی حدس‌های رمز عبور با شبیه‌سازی فرآیندهای انتخاب طبیعی استفاده می‌کند.

۳-۲۰-۲ دارایی‌های هدف

- **حساب‌های کاربری:**
  - حساب‌های ایمیل، پروفایل‌های شبکه‌های اجتماعی، پلتفرم‌های بانکداری آنلاین و برنامه‌های سازمانی که اطلاعات حساس را ذخیره می‌کنند.
- **خدمات ابری:**
  - برنامه‌ها و APIهای میزبانی‌شده در ابر که در معرض اینترنت قرار دارند و رمزهای عبور ضعیف می‌توانند منجر به دسترسی غیرمجاز شوند.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - لوازم خانگی هوشمند، دستگاه‌های پوشیدنی و سنسورهای صنعتی که اغلب با رمزهای عبور پیش‌فرض یا ضعیف پیکربندی شده‌اند.

- **سیستم‌های سازمانی:**
  - شبکه‌های شرکتی، سرورهای فایل و سیستم‌های پایگاه‌داده که داده‌های حیاتی کسب‌وکار را ذخیره می‌کنند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، سیستم‌های تأمین آب و شبکه‌های حمل‌ونقل که برای عملکرد به احراز هویت امن متکی هستند.

۲-۲۰-۴ آسیب‌پذیری‌ها

- **سیاست‌های ضعیف رمز عبور:**
  - عدم وجود الزامات پیچیدگی یا استفاده مجدد از رمزهای عبور در چندین حساب، حملات بروت فورس را آسان‌تر می‌کند.
- **محدودیت نرخ ناکافی:**
  - عدم وجود مکانیزم‌های محدودیت نرخ مؤثر، امکان انجام تلاش‌های سریع ورود را بدون ایجاد هشدار برای مهاجمان فراهم می‌کند.
- **الگوریتم‌های هش قدیمی:**
  - استفاده از الگوریتم‌های هش ضعیف یا نادرست، رمزهای عبور ذخیره‌شده را در معرض شکستن سریع‌تر قرار می‌دهد.
- **خطای انسانی:**
  - کاربرانی که رمزهای عبور قابل پیش‌بینی انتخاب می‌کنند یا احراز هویت چندعاملی را فعال نمی‌کنند، خطر به‌خطر افتادن را افزایش می‌دهند.
- **نقض داده‌ها:**
  - پایگاه‌داده‌های رمز عبور لورفته، داده‌های آموزشی ارزشمندی برای حملات مبتنی بر هوش مصنوعی فراهم می‌کنند و به مهاجمان امکان می‌دهند استراتژی‌های خود را بهبود بخشند.

۲-۲۰-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم مجموعه‌داده‌های بزرگی از رمزهای عبور لورفته، رفتار کاربران و الگوهای زبانی را با استفاده از ابزارهای مبتنی بر هوش مصنوعی جمع‌آوری می‌کند.

- **آموزش مدل:**
  - مهاجم مدل‌های یادگیری ماشین را بر روی داده‌های جمع‌آوری‌شده آموزش داده تا ساختارها، الگوها و ضعف‌های رایج رمز عبور را شناسایی کنند.
- **اجرای حمله:**
  - مهاجم مدل آموزش‌دیده را برای تولید و آزمایش حدس‌های رمز عبور در برابر سیستم هدف مستقر کرده و کاندیداهای با احتمال بالا را اولویت‌بندی می‌کند.
- **سازگاری:**
  - مهاجم به‌طور مداوم استراتژی حمله را بر اساس بازخورد تلاش‌های قبلی اصلاح می‌کند و با تغییرات در سیاست‌های رمز عبور یا اقدامات دفاعی سازگار می‌شود.
- **سوءاستفاده:**
  - مهاجم پس از یافتن یک رمز عبور معتبر، از آن برای دسترسی غیرمجاز به حساب یا سیستم هدف استفاده می‌کند.

۶-۲۰-۲ نشانه‌های احتمال نفوذ

- **تلاش‌های ناموفق ورود:**
  - افزایش ناگهانی در تلاش‌های ناموفق ورود به دنبال دسترسی موفق ممکن است نشان‌دهنده حمله بروت فورس باشد.
- **الگوهای فعالیت غیرعادی:**
  - ناهنجاری‌ها در زمان‌های ورود، مکان‌ها یا دستگاه‌هایی که با کاربران قانونی مرتبط نیستند.
- **استفاده افزایش‌یافته از منابع:**
  - مصرف بیشتر CPU یا حافظه به دلیل تلاش‌های مکرر ورود یا سایر فعالیت‌های خودکار.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول کاربران، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مرتبط با شکست‌های احراز هویت، قفل شدن حساب‌ها یا فعالیت غیرمنتظره.

۲-۲۰-۷ تأثیرات

- **دسترسی غیرمجاز:**
  - موفقیت در حمله بروت فورس منجر به دسترسی غیرمجاز به حساب‌ها یا سیستم‌های حساس می‌شود و امکان بهره‌برداری بیشتر را فراهم می‌کند.
- **نقض داده‌ها:**
  - سرقت داده‌های شخصی، مالی یا مالکیت فکری که منجر به ضررهای مالی قابل توجه یا آسیب به اعتبار می‌شود.
- **اختلال در عملیات:**
  - اختلال در سرویس‌ها یا سیستم‌های حیاتی کسب‌وکار به دلیل به‌خطر افتادن اعتبارنامه‌ها.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلال در خدمات منجر به ضررهای مالی می‌شود.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

۲-۲۰-۸ راه‌های مقابله

- **سیاست‌های قوی رمز عبور:**
  - الزامات پیچیدگی را اعمال کنید، رمزهای عبور رایج را ممنوع کنید و به‌روزرسانی منظم رمزهای عبور را اجباری کنید تا خطر حملات بروت فورس کاهش یابد.
- **احراز هویت چندعاملی:**
  - نیاز به مراحل تأیید اضافی فراتر از رمز عبور، مانند کدهای یک‌بارمصرف یا اسکن‌های بیومتریک، برای محافظت در برابر دسترسی غیرمجاز.
- **محدودیت نرخ:**
  - مکانیزم‌های محدودیت نرخ سخت‌گیرانه را پیاده‌سازی کنید تا تعداد تلاش‌های ورود در واحد زمان را محدود کنید و برای فعالیت مشکوک هشدار ایجاد کنید.
- **قفل شدن حساب‌ها:**
  - پس از تعداد مشخصی از تلاش‌های ناموفق ورود، حساب‌ها را به‌طور موقت غیرفعال کنید تا از حدس‌زنی مداوم جلوگیری شود.

- **هش کردن رمز عبور:**
  - از الگوریتم‌های هش قوی و نمک‌دار مانند Argon2، bcrypt یا PBKDF2 برای ذخیره ایمن رمزهای عبور و کند کردن تلاش‌های شکستن استفاده کنید.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای ورود غیرعادی یا فعالیت مشکوک نشان‌دهنده حملات بروت فورس استفاده کنید.
- **بررسی‌های منظم:**
  - ممیزی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد بهترین روش‌ها برای ایجاد رمزهای عبور قوی و تشخیص علائم خطر آموزش دهید.

۲-۲۱ جعل

حمله جعل[1] شامل جعل داده‌ها یا اعتبارنامه‌ها، مانند آدرس‌های IP، هدرهای ایمیل، آدرس‌های MAC یا اطلاعات بیومتریک، توسط یک مهاجم است که خود را به‌عنوان یک موجودیت قانونی جا می‌زند. در حمله جعل مبتنی بر هوش مصنوعی، از هوش مصنوعی برای افزایش پیچیدگی، دقت و مقیاس‌پذیری این تکنیک‌های فریب‌آمیز استفاده می‌شود. با استفاده از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند داده‌های جعلی بسیار متقاعدکننده ایجاد کنند، با اقدامات متقابل سازگار شوند و فرآیند فریب سیستم‌ها، کاربران یا شبکه‌ها را خودکار کنند.

حملات جعل مبتنی بر هوش مصنوعی خطرناک هستند زیرا می‌توانند مکانیزم‌های امنیتی سنتی را دور بزنند، از اعتماد انسان سوءاستفاده کنند و سیستم‌های پیچیده‌ای مانند تشخیص صدا، احراز هویت چهره یا پروتکل‌های شبکه را هدف قرار دهند.

۲-۲۱-۱ ویژگی‌های کلیدی

- **دقت بالا:**
  - هوش مصنوعی اطمینان می‌دهد که موجودیت‌های جعل‌شده بسیار شبیه به موجودیت‌های قانونی هستند و تشخیص آن‌ها دشوارتر می‌شود.

1. Spoofing Attack

- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند شناسایی اهداف، تولید داده‌های جعلی و اجرای حملات را خودکار می‌کنند و تلاش دستی را کاهش می‌دهند.
- **انعطاف‌پذیری:**
  - حملات جعل مبتنی بر هوش مصنوعی می‌توانند با تغییرات در دفاع سیستم یا رفتار کاربران سازگار شوند.
- **پنهان‌کاری:**
  - موجودیت‌های جعل‌شده به‌گونه‌ای طراحی شده‌اند که به‌راحتی با موجودیت‌های قانونی ترکیب شوند و از تشخیص توسط سیستم‌های تشخیص نفوذ (IDS) یا نرم‌افزارهای آنتی‌ویروس جلوگیری کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین موجودیت را به‌طور همزمان جعل کنند و دامنه و تأثیر حمله را افزایش دهند.

۲-۲۱-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های رفتارها یا ویژگی‌های موجودیت‌های قانونی آموزش می‌دهد تا جعل‌های دقیقی ایجاد کند.
  - یادگیری بدون نظارت: الگوهای فعالیت عادی را شناسایی می‌کند تا راهنمایی برای ایجاد جعل‌های متقاعدکننده بدون برچسب‌گذاری قبلی باشد.
  - یادگیری تقویتی: استراتژی‌های جعل را با پاداش دادن به فریب موفق و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در داده‌های تصویری، صوتی یا رفتاری را تحلیل می‌کنند تا جعل‌های واقع‌گرایانه تولید کنند.
  - شبکه‌های مولد تخاصمی: نسخه‌های مصنوعی اما واقع‌گرایانه از چهره‌ها، صداها یا الگوهای ترافیک شبکه ایجاد می‌کنند.

- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند ایمیل‌ها، لاگ‌های چت یا نام‌های دامنه، را تحلیل می‌کند تا سبک‌های ارتباطی و استفاده از زبان را برای فیشینگ یا جعل دامنه تکثیر کند.
- **بینایی کامپیوتر:**
  - در جعل بیومتریک برای ایجاد تصاویر یا ویدیوهای جعلی از چهره‌ها، عنبیه‌ها یا اثرانگشت‌ها استفاده می‌شود که می‌توانند سیستم‌های احراز هویت را فریب دهند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار کاربر یا سیستم استفاده می‌کند و امکان ایجاد جعل‌هایی را فراهم می‌کند که مانند نسخه‌های اصلی عمل می‌کنند.

۳-۲۱-۲ دارایی‌های هدف

- **سیستم‌های ایمیل:**
  - هدرهای ایمیل، آدرس‌های فرستنده یا نام‌های دامنه که برای کمپین‌های فیشینگ یا حملات جعل ایمیل کسب‌وکار (BEC)[1] جعل می‌شوند.
- **احراز هویت بیومتریک:**
  - سیستم‌های تشخیص چهره، اسکن اثرانگشت یا تأیید صدا که در برابر جعل‌های تولیدشده توسط هوش مصنوعی آسیب‌پذیر هستند.
- **دستگاه‌های موبایل:**
  - سیم‌کارت‌ها، شماره‌های تلفن یا شناسه‌های دستگاه که برای دسترسی غیرمجاز یا کلاهبرداری جعل می‌شوند.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - لوازم خانگی هوشمند، سنسورهای صنعتی یا دستگاه‌های پزشکی که برای احراز هویت به شناسه‌ها یا گواهی‌های منحصر به فرد متکی هستند.
- **سیستم‌های مالی:**
  - درگاه‌های پرداخت، برنامه‌های بانکی یا صرافی‌های ارز دیجیتال که برای تراکنش‌های کلاهبرداری هدف قرار می‌گیرند.

1. Business Email Compromise (BEC)

۴-۲۱-۲ آسیب‌پذیری‌ها

- **مکانیزم‌های احراز هویت ضعیف:**
  - احراز هویت تک‌عاملی یا الگوریتم‌های قابل پیش‌بینی تولید توکن، جعل موجودیت‌ها را برای مهاجمان آسان‌تر می‌کند.
- **مکانیزم‌های تأییدی ضعیف :**
  - عدم وجود فرآیندهای تأیید قوی برای داده‌های ورودی، مانند SPF/DKIM/DMARC برای ایمیل یا قفل کردن گواهی‌نامه برای HTTPS.
- **رمزنگاری ناکافی:**
  - پروتکل‌های رمزنگاری ضعیف یا نادرست، داده‌های حساس مورد نیاز برای جعل را در معرض دید قرار می‌دهند.
- **ثبت گزارش‌های ناکافی:**
  - روش‌های ضعیف ثبت گزارش‌ها، ردیابی و انتساب فعالیت‌های جعل را دشوار می‌کند.
- **خطای انسانی:**
  - تنظیمات نادرست فایروال‌ها، اعتبارنامه‌های مشترک یا سایر اشتباهات مدیران ممکن است به‌طور ناخواسته جعل را تسهیل کند.

۵-۲۱-۲ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره موجودیت هدف، از جمله رفتار، ویژگی‌ها و تعاملات آن استفاده می‌کند.
- **پیش‌پردازش:**
  - مهاجم داده‌های مرتبط، مانند متادیتای دستگاه، نمونه‌های بیومتریک یا الگوهای ارتباطی را با استفاده از ابزارهای شنود یا تکنیک‌های مهندسی اجتماعی ضبط می‌کند.
- **تولید جعل:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های جمع‌آوری‌شده و ایجاد یک جعل واقع‌گرایانه از موجودیت هدف مستقر می‌کند.
- **استقرار:**
  - مهاجم موجودیت جعل شده را در سیستم یا شبکه هدف معرفی کرده و اطمینان حاصل

می‌کند که به‌طور غیرقابل تشخیص از نسخه اصلی رفتار می‌کند.

- **سوءاستفاده:**
  - مهاجم از موجودیت جعل‌شده برای انجام اقدامات غیرمجاز، مانند دسترسی به منابع محدود، سرقت داده‌ها یا اختلال در عملیات استفاده می‌کند.
- **پایداری:**

مهاجم حضور موجودیت جعل‌شده را در طول زمان حفظ کرده و با تغییرات محیط سازگار شده تا از تشخیص جلوگیری کند.

۲-۲۱-۶ نشانه‌های احتمال نفوذ

- **عدم تطابق هدرها:**
  - ناسازگاری بین هدرهای ایمیل، آدرس‌های فرستنده یا نام‌های دامنه در حملات جعل ایمیل.
- **الگوهای ترافیکی غیرمنتظره:**
  - ناهنجاری‌ها در ترافیک شبکه، مانند آدرس‌های IP مبدأ غیرمعمول، پرس‌وجوهای DNS غیرمنتظره یا اندازه‌های بسته نامنظم.
- **افزایش تلاش‌های احراز هویت:**
  - افزایش ناگهانی در تلاش‌های ناموفق ورود به دنبال دسترسی موفق ممکن است نشان‌دهنده جعل باشد.
- **ورودی‌های غیرمنتظره در لاگ‌ها:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مرتبط با دستگاه‌های ناشناخته، دستورات غیرمعمول یا فعالیت غیرعادی.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول کاربران، مانند دسترسی به بخش‌های نامرتبط سیستم یا اجرای دستورات غیرمجاز.

۲-۲۱-۷ تأثیرات

- **سرقت هویت:**
  - دسترسی غیرمجاز به حساب‌های شخصی یا سازمانی، منجر به ضررهای مالی یا آسیب به اعتبار.

- **کلاهبرداری مالی:**
  - ابزارهای پرداخت جعل‌شده، مانند کارت‌های اعتباری یا حساب‌های بانکی، که برای تراکنش‌های کلاهبرداری استفاده می‌شوند.
- **اختلال در عملیات:**
  - دستگاه‌ها یا سیستم‌های جعل‌شده که باعث خرابی، اختلال یا از دست دادن خدمات حیاتی می‌شوند.
- **نقض داده‌ها:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالکیت فکری یا اسرار تجاری.
- **خطرات ایمنی:**
  - تهدید برای جان انسان‌ها در صورت دخالت موجودیت‌های جعل‌شده در زیرساخت‌های حیاتی، دستگاه‌های پزشکی یا سیستم‌های حمل‌ونقل.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، منجر به جریمه‌ها و پیامدهای قانونی می‌شود.
  - راه‌های مقابله
- **احراز هویت چندعاملی:**
  - نیاز به چندین شکل تأیید، مانند چیزی که می‌دانید (رمز عبور)، چیزی که دارید (توکن) و چیزی که هستید (بیومتریک).
- **احراز هویت دستگاه:**
  - مکانیزم‌های قوی شناسایی دستگاه را پیاده‌سازی کنید تا هر موجودیت را به‌طور منحصر به فرد شناسایی و احراز هویت کنید.
- **تحلیل‌های رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای ایجاد خطوط پایه و تشخیص انحرافات نشان‌دهنده کلون‌سازی استفاده کنید.
- **بررسی‌های منظم:**
  - ممیزی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **رمزنگاری و توکن‌سازی:**
  - داده‌های حساس را رمزنگاری کنید و به جای اعتبارنامه‌های ثابت از توکن‌های پویا استفاده

کنید تا خطر جعل کاهش یابد.

- **کنترل‌های دسترسی:**
  - اصول حداقل دسترسی را اجرا کنید و کنترل‌های دسترسی مبتنی بر نقش (RBAC)[1] را پیاده‌سازی کنید تا تأثیر حملات موفق را محدود کنید.
- **نظارت و ثبت لاگ:**
  - لاگ‌های دقیقی از فعالیت سیستم نگه دارید و با استفاده از ابزارهای نظارتی پیشرفته، رفتارهای مشکوک را نظارت کنید.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد بهترین روش‌ها برای ایمن‌سازی هویت‌های خود و تشخیص علائم جعل آموزش دهید.
- **پیاده‌سازی فناوری‌های ضد جعل:**
  - از فناوری‌هایی مانند DKIM، SPF و DMARC برای پیشگیری از جعل ایمیل استفاده کنید یا از تشخیص زنده‌بودن بیومتریک برای مقابله با تلاش‌های جعل احراز هویت استفاده کنید.

۲-۲۲ جعل بیومتریک

حمله جعل بیومتریک[2] شامل تلاش مهاجم برای فریب سیستم احراز هویت بیومتریک با ارائه داده‌های بیومتریک جعلی یا تکثیرشده، مانند اثر انگشت، ویژگی‌های چهره، ضبط صدا یا اسکن عنبیه است. در حمله جعل بیومتریک مبتنی بر هوش مصنوعی، از هوش مصنوعی برای افزایش واقع‌گرایی و دقت این جعل‌ها استفاده می‌شود و تشخیص آن‌ها را دشوارتر می‌کند. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند داده‌های بیومتریک مصنوعی بسیار متقاعدکننده‌ای ایجاد کنند که از اقدامات ضد جعل سنتی عبور می‌کنند.

این نوع حمله خطرناک است زیرا اعتماد به سیستم‌های بیومتریک را که به‌طور فزاینده‌ای برای دسترسی امن به سیستم‌ها، دستگاه‌ها و خدمات حساس استفاده می‌شوند، تضعیف می‌کند.

۲-۲۲-۱ ویژگی‌های کلیدی

- **تکثیر با کیفیت بالا:**
  - هوش مصنوعی اطمینان می‌دهد که داده‌های بیومتریک جعلی به‌طور نزدیکی ویژگی‌های دنیای واقعی را تقلید می‌کنند و احتمال فریب موفقیت‌آمیز را افزایش می‌دهند.

1. Role-Based Access Control (RBAC)
2. Biometric Spoofing attack

- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند تولید و استقرار داده‌های بیومتریک جعلی را خودکار می‌کنند و تلاش دستی را کاهش می‌دهند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در سیستم‌های بیومتریک، تکنیک‌های ضد جعل یا شرایط محیطی سازگار شوند.
- **پنهان‌کاری:**
  - داده‌های بیومتریک جعلی به‌گونه‌ای طراحی شده‌اند که به‌طور یکپارچه با ورودی‌های قانونی ترکیب شوند و از تشخیص توسط مکانیزم‌های تشخیص زنده‌بودن اجتناب کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین سیستم یا کاربر را به‌طور همزمان هدف‌گیری کنند، که این امر دامنه و تأثیر حمله را افزایش می‌دهد.

۲-۲۲-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌های داده‌ای از نمونه‌های بیومتریک قانونی آموزش می‌دهد تا ویژگی‌های متمایزکننده آن‌ها را یاد بگیرد و جعل‌های واقع‌گرایانه ایجاد کند.
  - یادگیری بدون نظارت: الگوهای داده‌های بیومتریک را بدون برچسب‌گذاری قبلی شناسایی می‌کند و امکان ایجاد جعل‌های متقاعدکننده را حتی زمانی که داده‌های برچسب‌خورده محدود هستند، فراهم می‌کند.
  - یادگیری تقویتی: استراتژی‌های جعل را با پاداش دادن به فریب موفقیت‌آمیز و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های مولد تخاصمی، الگوهای پیچیده در داده‌های بیومتریک را تحلیل می‌کنند تا نسخه‌های بسیار واقع‌گرایانه تولید کنند.
  - شبکه‌های مولد تخاصمی: داده‌های بیومتریک مصنوعی اما زنده‌مانند، مانند چهره‌های جعلی، صداها یا اثر انگشت‌هایی ایجاد می‌کنند که حتی سیستم‌های بیومتریک پیشرفته

را فریب می‌دهند.

- **پردازش زبان طبیعی:**
  - الگوهای زبانی در ضبط‌های صوتی را تحلیل می‌کند تا ویژگی‌های گفتاری، لحن و آهنگ صدا را برای جعل بیومتریک مبتنی بر صدا تقلید کند.
- **بینایی کامپیوتر:**
  - در جعل تشخیص چهره استفاده می‌شود تا تصاویر یا ویدیوهای جعلی از چهره‌هایی ایجاد کند که ظاهر و حرکات کاربران واقعی را تقلید می‌کنند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار کاربر استفاده می‌کند و به مهاجمان امکان می‌دهد نه تنها ویژگی‌های بیومتریک ایستا بلکه ویژگی‌های پویا مانند الگوهای تایپ یا تحلیل راه رفتن را نیز تقلید کنند.

۳-۲۲-۲ دارایی‌های هدف

- **سیستم‌های تشخیص چهره:**
  - دوربین‌های امنیتی، دستگاه‌های همراه و سیستم‌های کنترل دسترسی که از احراز هویت مبتنی بر چهره استفاده می‌کنند.
- **اسکنرهای اثر انگشت:**
  - دستگاه‌ها و سیستم‌هایی که از سنسورهای اثر انگشت برای دسترسی امن استفاده می‌کنند، مانند تلفن‌های هوشمند، لپ‌تاپ‌ها یا نقاط ورود فیزیکی.
- **سیستم‌های تشخیص صدا:**
  - دستیارهای مجازی، مراکز تماس و خدمات مالی که از احراز هویت مبتنی بر صدا استفاده می‌کنند.
- **اسکنرهای عنبیه:**
  - سیستم‌های امنیتی بالا که به شناسایی مبتنی بر عنبیه نیاز دارند، مانند کنترل مرزها یا تأسیسات شرکتی.
- **سیستم‌های بهداشتی:**
  - دستگاه‌های پزشکی یا پلتفرم‌هایی که از بیومتریک برای تأیید بیمار یا دسترسی به سوابق استفاده می‌کنند.

۲-۲۲-۴ آسیب‌پذیری‌ها

- **مکانیزم‌های ضعیف ضد جعل:**
  - عدم وجود تکنیک‌های قوی تشخیص زنده‌بودن یا ضد جعل، سیستم‌های بیومتریک را در برابر جعل آسیب‌پذیر می‌کند.
- **عدم رمزنگاری:**
  - ذخیره‌سازی یا انتقال داده‌های بیومتریک بدون رمزنگاری، آن‌ها را در معرض رهگیری و تکثیر قرار می‌دهد.
- **اعتبارسنجی ناکافی:**
  - عدم اعتبارسنجی صحیح ورودی‌های بیومتریک، اجازه می‌دهد داده‌های جعلی بدون تشخیص عبور کنند.
- **خطای انسانی:**
  - سیستم‌های نادرست پیکربندی‌شده یا پیاده‌سازی نادرست فناوری‌های بیومتریک، به‌طور ناخواسته جعل را تسهیل می‌کنند.
- **نشت داده:**
  - داده‌های بیومتریک نشت‌یافته، مواد آموزشی ارزشمندی برای حملات جعل مبتنی بر هوش مصنوعی فراهم می‌کنند.

۲-۲۲-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری داده‌های بیومتریک از منابع عمومی، پایگاه‌های داده نشت‌یافته یا نظارت پنهان استفاده می‌کند.
- **آموزش مدل:**
  - مدل‌های یادگیری ماشین را بر روی داده‌های جمع‌آوری‌شده آموزش می‌دهد تا نمونه‌های بیومتریک جعلی واقع‌گرایانه ایجاد کند.
- **تولید جعل:**
  - از شبکه‌های مولد تخاصمی یا سایر مدل‌های مولد برای ایجاد داده‌های بیومتریک مصنوعی اما متقاعدکننده، مانند چهره‌های جعلی، صداها یا اثر انگشت‌ها استفاده می‌کند.

- **استقرار:**
  - داده‌های بیومتریک جعلی را به سیستم هدف ارائه می‌دهد و اطمینان می‌یابد که به‌طور نزدیکی ورودی قانونی را تقلید می‌کند تا از احراز هویت عبور کند.
- **سوءاستفاده:**
  - دسترسی غیرمجاز به سیستم یا خدمات هدف را به‌دست می‌آورد و امکان سوءاستفاده بیشتر یا سرقت اطلاعات حساس را فراهم می‌کند.
- **پایداری:**
  - دسترسی را در طول زمان حفظ می‌کند و به‌طور مداوم داده‌های جعلی را بهبود می‌بخشد تا با هرگونه به‌روزرسانی در سیستم بیومتریک سازگار شود.

۲-۲۲-۶ نشانه‌های احتمال نفوذ

- **الگوهای دسترسی غیرمعمول:**
  - ناهنجاری‌ها در زمان‌ها، مکان‌ها یا دستگاه‌های ورود که با کاربران قانونی مرتبط نیستند.
- **افزایش شکست‌های احراز هویت:**
  - افزایش ناگهانی در تلاش‌های ناموفق احراز هویت که با دسترسی موفقیت‌آمیز همراه است، می‌تواند نشان‌دهنده جعل باشد.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر ناسازگار است، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به شکست‌های احراز هویت، قفل شدن حساب‌ها یا فعالیت‌های غیرمنتظره.
- **سرنخ‌های بصری/فیزیکی:**
  - نشانه‌های دستکاری در سنسورهای بیومتریک، مانند خراش‌ها، باقی‌مانده‌ها یا سایش غیرمعمول.

۲-۲۲-۷ تأثیرات

- **دسترسی غیرمجاز:**
  - جعل موفقیت‌آمیز منجر به دسترسی غیرمجاز به حساب‌ها، سیستم‌ها یا فضاهای فیزیکی

حساس می‌شود.

- **نشت داده:**
  - سرقت داده‌های شخصی، مالی یا مالکیت فکری که منجر به خسارات مالی قابل توجه یا آسیب به اعتبار می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل نفوذ به اعتبارنامه‌ها.
- **خسارات مالی:**
  - تراکنش‌های غیرمجاز، کلاهبرداری یا اختلال در خدمات که منجر به خسارات مالی می‌شود.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

۸-۲۲-۲ راه‌های مقابله

- **تشخیص زنده‌بودن:**
  - تکنیک‌های پیشرفته تشخیص زنده‌بودن، مانند تحلیل بافت، تشخیص حرکت یا تصویربرداری حرارتی را پیاده‌سازی کنید تا اصالت ورودی‌های بیومتریک تأیید شود.
- **احراز هویت چندعاملی:**
  - احراز هویت بیومتریک را با عوامل اضافی مانند رمز عبور یا توکن ترکیب کنید تا وابستگی به یک روش واحد کاهش یابد.
- **رمزنگاری:**
  - داده‌های بیومتریک را در هنگام ذخیره‌سازی و انتقال رمزنگاری کنید تا از رهگیری و تکثیر جلوگیری شود.
- **به‌روزرسانی‌های منظم:**
  - سیستم‌های بیومتریک و مکانیزم‌های ضد جعل را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده برطرف شوند.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص ناهنجاری‌ها در رفتار کاربر یا الگوهای ورودی بیومتریک که نشان‌دهنده جعل هستند، استفاده کنید.

- **آموزش آگاهی امنیتی:**
  - کارکنان و کاربران را در مورد تشخیص علائم جعل بیومتریک و رعایت بهداشت امنیت سایبری آموزش دهید.
- **مقاومت در برابر دستکاری:**
  - سنسورهای بیومتریک را با ویژگی‌های مقاوم در برابر دستکاری طراحی کنید تا از دستکاری فیزیکی یا تلاش‌های جعل جلوگیری شود.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.

۲-۲۳ سرقت هویت

سرقت هویت[1] شامل تلاش مهاجم برای جعل یا استفاده از هویت یک فرد، معمولاً برای کسب سود مالی، کلاهبرداری یا سایر اهداف مخرب است. در حمله سرقت هویت مبتنی بر هوش مصنوعی، از هوش مصنوعی برای بهبود و خودکارسازی فرآیند سرقت، ترکیب و سوءاستفاده از اطلاعات شخصی استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند حجم عظیمی از داده‌ها را از شبکه‌های اجتماعی، سوابق عمومی و نشت‌های داده تحلیل کنند تا هویت‌های جعلی بسیار متقاعدکننده ایجاد کنند یا هویت‌های واقعی را با دقت بالا تکثیر کنند.

این نوع حمله می‌تواند از مکانیزم‌های احراز هویت سنتی عبور کند، از روان‌شناسی انسان از طریق مهندسی اجتماعی سوءاستفاده کند و به‌طور پویا با تغییرات در اقدامات امنیتی سازگار شود.

۲-۲۳-۱ ویژگی‌های کلیدی

- **دقت مبتنی بر داده:**
  - هوش مصنوعی مجموعه‌های بزرگ داده‌های شخصی، مانند نام‌ها، آدرس‌ها، شماره‌های تأمین اجتماعی و داده‌های بیومتریک را تحلیل می‌کند تا هویت‌های جعلی واقع‌گرایانه ایجاد کند یا هویت‌های واقعی را تکثیر کند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند جمع‌آوری، ترکیب و سوءاستفاده از هویت‌های سرقت‌شده را خودکار می‌کنند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهند.

1. Identity Theft

- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در اقدامات امنیتی، رفتار کاربر یا فناوری‌های دفاعی سازگار شوند.
- **پنهان‌کاری:**
  - با تقلید از فعالیت‌های قانونی یا ترکیب شدن با نویز پس‌زمینه، حملات سرقت هویت مبتنی بر هوش مصنوعی می‌توانند برای مدت‌های طولانی کشف نشوند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین فرد یا سیستم را به‌طور همزمان هدف‌گیری کنند، که این امر دامنه و تأثیر حمله را افزایش می‌دهد.

۲-۲۳-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌های داده‌ای از هویت‌های شناخته‌شده آموزش می‌دهد تا الگوها را یاد بگیرد و هویت‌های مصنوعی واقع‌گرایانه ایجاد کند.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در داده‌های بدون برچسب شناسایی می‌کند تا روابط پنهان یا فعالیت‌های مشکوک را کشف کند.
  - یادگیری تقویتی: فرآیند سرقت هویت را با پاداش دادن به جعل‌های موفقیت‌آمیز و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در تصاویر، متن یا داده‌های رفتاری را تحلیل می‌کنند تا هویت‌های مصنوعی زنده‌مانند ایجاد کنند.
  - شبکه‌های مولد تخاصمی: هویت‌های مصنوعی اما واقع‌گرایانه، از جمله چهره‌ها، صداها یا اسناد ایجاد می‌کنند که می‌توانند سیستم‌های تأیید هویت را فریب دهند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند پست‌های شبکه‌های اجتماعی، ایمیل‌ها یا سوابق عمومی را تحلیل می‌کند تا جزئیات شخصی را استخراج کند و پیام‌های فیشینگ متقاعدکننده یا ارتباطات کلاهبرداری ایجاد کند.

- **بینایی کامپیوتر:**
  - در جعل تشخیص چهره استفاده می‌شود تا تصاویر یا ویدیوهای جعلی از افراد برای دور زدن تأیید هویت ایجاد کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر استفاده می‌کند و به مهاجمان امکان می‌دهد تا کاربران قانونی را به‌طور متقاعدکننده جعل کنند.

۳-۲۳-۲ دارایی‌های هدف

- **اطلاعات شخصی:**
  - نام‌ها، آدرس‌ها، شماره‌های تأمین اجتماعی، جزئیات حساب بانکی، آدرس‌های ایمیل و رمزهای عبور.
- **داده‌های بیومتریک:**
  - اثر انگشت، اسکن‌های چهره، ضبط‌های صوتی و الگوهای عنبیه که برای تأیید هویت استفاده می‌شوند.
- **سیستم‌های مالی:**
  - بانک‌ها، شرکت‌های کارت اعتباری و پلتفرم‌های پرداخت آنلاین که داده‌های مالی حساس را ذخیره می‌کنند.
- **خدمات دولتی:**
  - سازمان‌های مالیاتی، ارائه‌دهندگان خدمات بهداشتی و سایر مؤسساتی که برای دسترسی به تأیید هویت نیاز دارند.
- **سیستم‌های سازمانی:**
  - شبکه‌های شرکتی، سرورهای فایل و سیستم‌های پایگاه داده که داده‌های حیاتی کسب‌وکار مرتبط با هویت کارمندان را ذخیره می‌کنند.

۴-۲۳-۲ آسیب‌پذیری‌ها

- **نشت داده:**
  - پایگاه‌های داده نشت‌یافته حاوی اطلاعات شخصی، مواد ارزشمندی برای حملات سرقت هویت مبتنی بر هوش مصنوعی فراهم می‌کنند.

- **مکانیزم‌های احراز هویت ضعیف:**
  - احراز هویت تک‌عاملی یا الگوریتم‌های تولید توکن قابل پیش‌بینی، سرقت هویت را برای مهاجمان آسان‌تر می‌کند.
- **خطای انسانی:**
  - کاربرانی که اطلاعات شخصی خود را به‌طور بیش‌ازحد در شبکه‌های اجتماعی به اشتراک می‌گذارند یا در دام کلاهبرداری‌های فیشینگ می‌افتند، به‌طور ناخواسته هویت خود را در معرض خطر قرار می‌دهند.
- **کنترل‌های حریم خصوصی ناکافی:**
  - عدم وجود رمزنگاری قوی، ناشناس‌سازی یا کنترل‌های دسترسی، داده‌های حساس را در معرض دسترسی غیرمجاز قرار می‌دهد.
- **سیستم‌های قدیمی:**
  - استفاده از نرم‌افزارها، فرم‌ورها یا پروتکل‌های قدیمی که حاوی آسیب‌پذیری‌های شناخته‌شده قابل سوءاستفاده توسط سارقان هویت هستند.

۵-۲۳-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات شخصی از منابع عمومی، پایگاه‌های داده نشت‌یافته یا نظارت پنهان استفاده می‌کند.
- **ترکیب:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های جمع‌آوری‌شده مستقر می‌کند و هویت‌های مصنوعی اما واقع‌گرایانه ایجاد می‌کند یا هویت‌های واقعی را تکثیر می‌کند.
- **سوءاستفاده:**
  - مهاجم از هویت‌های ترکیب‌شده یا تکثیرشده برای ارتکاب کلاهبرداری، دسترسی غیرمجاز یا انجام سایر فعالیت‌های مخرب استفاده می‌کند.
- **پایداری:**
  - مهاجم دسترسی به حساب‌ها یا سیستم‌های به‌خطر افتاده را با بهبود مداوم هویت‌های سرقت‌شده یا مصنوعی برای سازگاری با تغییرات در اقدامات امنیتی حفظ می‌کند.
- **پوشاندن ردپا:**
  - مهاجم لاگ‌ها را تغییر می‌دهد، ردپاها را پاک می‌کند یا از تکنیک‌های دیگر برای اجتناب از

تشخیص و حفظ پنهان‌کاری استفاده می‌کند.

۲-۲۳-۶ نشانه‌های احتمال نفوذ

- **فعالیت غیرعادی حساب:**
  - ناهنجاری‌ها در زمان‌ها، مکان‌ها یا دستگاه‌های ورود که با کاربران قانونی مرتبط نیستند.
- **تراکنش‌های غیرمنتظره:**
  - تراکنش‌های مالی مشکوک، مانند خریدها یا برداشت‌ها، که از حساب‌های ناآشنا انجام می‌شوند.
- **افزایش تلاش‌های احراز هویت:**
  - افزایش ناگهانی در تلاش‌های ناموفق ورود که با دسترسی موفقیت‌آمیز همراه است، می‌تواند نشان‌دهنده سرقت هویت باشد.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر ناسازگار است، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ایمیل‌های فیشینگ:**
  - ایمیل‌های مشکوک حاوی پیوست‌ها یا لینک‌های مخرب که برای سرقت اطلاعات شخصی طراحی شده‌اند.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به احراز هویت ناموفق، قفل شدن حساب‌ها یا فعالیت‌های غیرمنتظره.

۲-۲۳-۷ تأثیرات

- **خسارات مالی:**
  - تراکنش‌های غیرمجاز، وام‌ها یا درخواست‌های کارت اعتباری که منجر به خسارات مالی قابل توجهی می‌شوند.
- **آسیب به اعتبار:**
  - از دست دادن اعتماد مشتریان، شرکا یا ذینفعان پس از نفوذ یا سوءاستفاده از هویت‌های سرقت‌شده.
- **پیامدهای قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و اقدامات قانونی می‌شود.

- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل به‌خطر افتادن اعتبارنامه‌ها یا سوءاستفاده از هویت‌ها.
- **استرس عاطفی:**
  - قربانیان سرقت هویت ممکن است استرس، اضطراب یا سایر اثرات روانی ناشی از مواجهه با عواقب آن را تجربه کنند.

۸-۲۳-۲ راه‌های مقابله

- **احراز هویت چندعاملی:**
  - مراحل تأیید اضافی مانند کدهای یک‌بارمصرف یا اسکن‌های بیومتریک را فراتر از رمز عبور الزامی کنید تا از دسترسی غیرمجاز جلوگیری شود.
- **رمزنگاری:**
  - داده‌های حساس را در هنگام ذخیره‌سازی و انتقال رمزنگاری کنید تا از رهگیری و سوءاستفاده جلوگیری شود.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرم‌ورها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده را برطرف کنید.
- **آموزش آگاهی امنیتی:**
  - کارکنان و کاربران را در مورد تشخیص تلاش‌های فیشینگ، اجتناب از دانلودهای مخرب و رعایت بهداشت امنیت سایبری آموزش دهید.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی ورود یا فعالیت‌های مشکوک که نشان‌دهنده سرقت هویت هستند، استفاده کنید.
- **کاهش داده:**
  - فقط حداقل داده‌های لازم را جمع‌آوری و ذخیره کنید تا تأثیر احتمالی نفوذ کاهش یابد.
- **تأیید هویت:**
  - مکانیزم‌های پیشرفته تأیید هویت، مانند تشخیص زنده‌بودن یا بررسی‌های چندلایه را پیاده‌سازی کنید تا اصالت را تضمین کنید.

- **حسابرسی‌های منظم:**
  - حسابرسی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **سیستم‌های تشخیص کلاهبرداری:**
  - ابزارهای تشخیص کلاهبرداری مبتنی بر هوش مصنوعی را مستقر کنید تا تراکنش‌ها یا فعالیت‌های حساب مشکوک را نظارت کنند.

۲-۲۴ تزریق اسکریپت‌های بین‌سایتی

حمله‌های تزریق اسکریپت‌های بین‌سایتی نوعی آسیب‌پذیری امنیتی در وب هستند که به مهاجمان اجازه می‌دهند اسکریپت‌های مخرب را در وب‌سایت‌ها یا برنامه‌های وب قانونی که توسط کاربران دیگر مشاهده می‌شوند، تزریق کنند. در یک حمله اسکریپت‌گذاری بین‌سایتی مبتنی بر هوش مصنوعی، از هوش مصنوعی برای خودکارسازی و بهبود فرآیند شناسایی آسیب‌پذیری‌ها، ایجاد payloadها و دور زدن مکانیزم‌های تشخیص استفاده می‌شود. با استفاده از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند مجموعه‌داده‌های بزرگی از رفتارهای برنامه‌های وب را تحلیل کنند، نقاط تزریق بالقوه را پیش‌بینی کنند و payloadهای بسیار مؤثر و متناسب با اهداف خاص ایجاد کنند.

این نوع حمله از آن جهت مهم و خطرناک است که می‌تواند دفاع‌های سنتی مانند پاک‌سازی ورودی‌ها، سیاست‌های امنیتی محتوا (CSPs)[1] یا فایروال‌های برنامه‌های وب (WAFs)[2] را دور بزنند. حملات اسکریپت‌گذاری بین‌سایتی مبتنی بر هوش مصنوعی از خطاهای انسانی در کدنویسی، مکانیزم‌های اعتبارسنجی ضعیف و فناوری‌های وب در حال تکامل سوءاستفاده می‌کنند تا جلسات کاربران را به خطر بیندازند، داده‌های حساس را سرقت کنند یا بدافزار منتشر کنند.

۲-۲۴-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند اسکن برای آسیب‌پذیری‌ها، ایجاد payloadها و اجرای حملات را خودکار می‌کند، که تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا نقاط تزریق خاص را با دقت بالا شناسایی کنند و حداکثر تأثیر را داشته باشند.

1. Content Security Policy (CSP)
2. Web Application Firewall (WAFs)

- **انطباق‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در منطق برنامه‌های وب، اقدامات دفاعی یا شرایط محیطی سازگار شوند.
- **پنهان‌کاری:**
  - با تقلید از ترافیک قانونی یا رمزنگاری payloadها تا زمان اجرا، هوش مصنوعی اطمینان می‌دهد که اسکریپت‌های مخرب شناسایی نشوند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین برنامه وب را به‌طور همزمان اسکن و مورد سوءاستفاده قرار دهند، که پتانسیل آسیب را افزایش می‌دهد.

۲-۲۴-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های برچسب‌دار از آسیب‌پذیری‌های شناخته‌شده اسکریپت‌گذاری بین‌سایتی و payloadهای موفق آموزش می‌دهد تا الگوها را یاد بگیرد و موارد جدید را شناسایی کند.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در رفتار برنامه‌های وب را شناسایی می‌کند تا آسیب‌پذیری‌های پنهان را آشکار کند.
  - یادگیری تقویتی: تولید payloadها را با پاداش دادن به سوءاستفاده موفق و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، الگوهای پیچیده در ورودی‌ها، خروجی‌ها و تعاملات برنامه‌های وب را تحلیل می‌کنند تا استراتژی‌های حمله را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: payloadهای مصنوعی اما واقع‌گرایانه ایجاد می‌کنند که برای دور زدن ابزارهای تحلیل ایستا یا دفاع‌های مبتنی بر امضا طراحی شده‌اند.
- **تکنیک‌های فازی:**
  - هوش مصنوعی را با روش‌های فازی ترکیب می‌کند تا برنامه‌های وب را به‌طور سیستماتیک برای رفتارهای غیرمنتظره یا آسیب‌پذیری‌ها تست کند.

- **پردازش زبان طبیعی:**
  - داده‌های متنی مانند JavaScript، HTML یا محتوای تولیدشده توسط کاربر را تحلیل می‌کند تا زمینه‌های بیشتری درباره نقاط تزریق بالقوه استنباط کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا اسکریپت‌هایی ایجاد کند که به‌طور یکپارچه با فعالیت قانونی ترکیب شوند.

۳-۲۴-۲ دارایی‌های هدف

- **برنامه‌های وب:**
  - پلتفرم‌های تجارت الکترونیک، سایت‌های شبکه‌های اجتماعی یا خدمات مالی که در آن‌ها اسکریپت‌های تزریق‌شده می‌توانند کوکی‌های جلسه، اعتبار یا داده‌های حساس را سرقت کنند.
- **سیستم‌های سازمانی:**
  - اینترانت‌های شرکتی، سیستم‌های مدیریت ارتباط با مشتری (CRM)[1] یا پورتال‌های کارمندی که در برابر تزریق اسکریپت آسیب‌پذیر هستند.
- **جلسات کاربران:**
  - جلسات کاربران به‌خطرافتاده که منجر به دسترسی غیرمجاز، سرقت داده یا سوءاستفاده بیشتر می‌شود.
- **ادغام‌های شخص ثالث:**
  - افزونه‌ها، ویجت‌ها یا APIهایی که در سیستم‌های بزرگتر ادغام شده‌اند و نقاط ورودی برای حملات اسکریپت‌گذاری بین‌سایتی فراهم می‌کنند.
- **زیرساخت‌های حیاتی:**
  - رابط‌های وب‌محور که شبکه‌های برق، سیستم‌های بهداشتی یا شبکه‌های حمل‌ونقل را کنترل می‌کنند و در آن‌ها اسکریپت‌گذاری بین‌سایتی می‌تواند آسیب‌های قابل توجهی ایجاد کند.

۴-۲۴-۲ آسیب‌پذیری‌ها

- **اعتبارسنجی ضعیف ورودی‌ها:**
  - فقدان اعتبارسنجی یا پاک‌سازی قوی ورودی‌های کاربر، اجازه می‌دهد اسکریپت‌های مخرب

1. CRM

بدون بررسی عبور کنند.

- **رمزگذاری نامناسب خروجی‌ها:**
  - عدم رمزگذاری صحیح خروجی‌ها، برنامه‌های وب را در برابر حملات اسکریپت‌گذاری بین‌سایتی بازتابی یا ذخیره‌شده آسیب‌پذیر می‌کند.
- **نرم‌افزارهای قدیمی:**
  - استفاده از کتابخانه‌ها، فریم‌ورک‌ها یا سیستم‌های عامل قدیمی که حاوی آسیب‌پذیری‌های شناخته‌شده قابل سوءاستفاده از طریق اسکریپت‌گذاری بین‌سایتی هستند.
- **تست امنیتی ناکافی:**
  - تست ناکافی برای اسکریپت‌گذاری بین‌سایتی در طول توسعه، برنامه‌ها را در معرض سوءاستفاده قرار می‌دهد.
- **خطای انسانی:**
  - توسعه‌دهندگان ممکن است از طریق روش‌های ضعیف کدنویسی یا پیاده‌سازی نادرست اقدامات امنیتی، آسیب‌پذیری‌ها را ایجاد کنند.

۵-۲۴-۲ سناریوی تهدید

- **شناسایی:**
  - از ابزارهای مبتنی بر هوش مصنوعی برای نقشه‌برداری از برنامه وب هدف استفاده کنید و نقاط تزریق بالقوه مانند نوارهای جستجو، بخش‌های نظرات یا فیلدهای فرم را شناسایی کنید.
- **اسکن آسیب‌پذیری:**
  - مدل‌های یادگیری ماشین را برای تحلیل برنامه وب از نظر ضعف‌هایی مانند اعتبارسنجی نادرست ورودی‌ها یا رمزگذاری ضعیف خروجی‌ها مستقر کنید.
- **تولید payload:**
  - از یادگیری عمیق یا شبکه‌های مولد تخاصمی برای ایجاد payloadهای سفارشی که متناسب با آسیب‌پذیری‌های شناسایی‌شده هستند، استفاده کنید.
- **اجرای تزریق:**
  - payload طراحی‌شده را از طریق نقطه تزریق انتخاب‌شده به برنامه هدف تحویل دهید و اطمینان حاصل کنید که با ترافیک عادی ترکیب می‌شود.

- **سوءاستفاده:**
  - اسکریپت تزریق‌شده را اجرا کنید تا به نتیجه مطلوب برسید، مانند سرقت کوکی‌های جلسه، تغییر مسیر کاربران یا انتشار بدافزار.
- **پایداری:**
  - لاگ‌ها را تغییر دهید، ردپاها را پاک کنید یا از تکنیک‌های دیگر برای جلوگیری از تشخیص و حفظ دسترسی استفاده کنید.

۲-۲۴-۶ نشانه‌های احتمال نفوذ

- **اسکریپت‌های غیرمعمول:**
  - ناهنجاری‌ها در کد JavaScript یا HTML جاسازی‌شده در صفحات وب که با عملکرد قانونی مرتبط نیستند.
- **تغییر مسیرهای غیرمنتظره:**
  - کاربران پس از تعامل با برنامه به‌خطرافتاده به وب‌سایت‌های ناشناخته یا مشکوک هدایت می‌شوند.
- **افزایش تلاش‌های احراز هویت:**
  - افزایش ناگهانی در تلاش‌های ناموفق ورود به سیستم که با دسترسی موفق دنبال می‌شود، ممکن است نشان‌دهنده سرقت اعتبار از طریق اسکریپت‌گذاری بین‌سایتی باشد.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربران ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سرور مربوط به ورودی‌های غیرمنتظره، خطاها یا فعالیت غیرعادی.

۲-۲۴-۷ تأثیرات

- **نشت داده‌ها:**
  - سرقت داده‌های شخصی، مالی یا مالکیت فکری که منجر به ضررهای مالی قابل توجه یا آسیب‌های اعتباری می‌شود.

- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، فعالیت‌های کلاهبرداری یا اختلال در خدمات که باعث آسیب‌های مالی می‌شوند.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به‌دلیل برنامه‌های وب به‌خطرافتاده.
- **آسیب به شهرت:**
  - از دست دادن اعتماد مشتریان و ارزش برند پس از یک نفوذ یا سوءاستفاده از داده‌های سرقت‌شده.
- **توزیع بدافزار:**
  - استفاده از برنامه‌های وب به‌خطرافتاده برای انتشار بدافزار یا باج‌افزار به کاربران بی‌اطلاع.

۸-۲۴-۲ راه‌های مقابله

- **اعتبارسنجی ورودی‌ها:**
  - قوانین سخت‌گیرانه اعتبارسنجی ورودی‌ها را پیاده‌سازی کنید تا اطمینان حاصل شود که تمام ورودی‌های کاربر با فرمت‌ها و مقادیر مورد انتظار مطابقت دارند.
- **رمزگذاری خروجی‌ها:**
  - خروجی‌ها را رمزگذاری کنید تا هرگونه کاراکتر بالقوه مضر قبل از رندر شدن در رابط کاربری خنثی شود.
- **سیاست امنیتی محتوا (CSP):**
  - هدرهای CSP را اعمال کنید تا منابع اسکریپت‌های قابل اجرا را محدود کرده و از اجرای کد غیرمجاز جلوگیری کنید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، کتابخانه‌ها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده را برطرف کنید.
- **تست امنیتی:**
  - تست نفوذ، ارزیابی آسیب‌پذیری و بررسی کد را به‌طور منظم انجام دهید تا ضعف‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.

- **نظارت رفتاری:**
  - از تحلیل رفتاری مبتنی بر هوش مصنوعی استفاده کنید تا الگوهای فعالیت غیرعادی که نشان‌دهنده تلاش‌های اسکریپت‌گذاری بین‌سایتی هستند، شناسایی شوند.
- **آموزش و آگاهی:**
  - توسعه‌دهندگان و اپراتورها را در مورد روش‌های کدنویسی ایمن و اهمیت اعتبارسنجی و پاک‌سازی ورودی‌ها آموزش دهید.
- **فایروال برنامه وب:**
  - راه‌حل‌های پیشرفته فایروال برنامه وب را مستقر کنید که قادر به شناسایی و مسدود کردن حملات اسکریپت‌گذاری بین‌سایتی در زمان واقعی هستند.
- **مدیریت جلسات:**
  - روش‌های مدیریت جلسات قوی مانند بازتولید شناسه‌های جلسه پس از احراز هویت و اعمال زمان‌بندی کوتاه را پیاده‌سازی کنید.
- **هوشمندسازی تهدیدات پیش‌گیرانه:**
  - از پلتفرم‌های هوشمند مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور اسکریپت‌گذاری بین‌سایتی و تهدیدات تطبیقی مطلع شوید.

۲-۲۵ تزریق SQL

حمله تزریق SQL[1] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای خودکارسازی و بهبود فرآیند سوءاستفاده از آسیب‌پذیری‌ها در برنامه‌های وب یا پایگاه‌داده‌هایی است که اجازه اجرای پرس‌وجوهای SQL مخرب را می‌دهند. با استفاده از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند مجموعه‌داده‌های بزرگی از رفتارهای برنامه‌ها را تحلیل کنند، نقاط تزریق بالقوه را پیش‌بینی کنند و payloadهای بسیار مؤثر و متناسب با اهداف خاص ایجاد کنند. این نوع حمله به‌خصوص خطرناک است زیرا می‌تواند دفاع‌های سنتی مانند اعتبارسنجی ورودی‌ها، دستورات آماده‌شده[2] یا فایروال‌های برنامه‌های وب را دور بزند.

حملات تزریق SQL مبتنی بر هوش مصنوعی از خطاهای انسانی در کدنویسی، مکانیزم‌های اعتبارسنجی ضعیف و فناوری‌های پایگاه‌داده در حال تکامل سوءاستفاده می‌کنند تا داده‌های حساس را به خطر بیندازند، رکوردها را دستکاری کنند یا دسترسی غیرمجاز به سیستم‌های بک‌اند به دست آورند.

1. SQL Injection
2. prepared statements

۲-۲۵-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند اسکن برای آسیب‌پذیری‌ها، ایجاد payloadها و اجرای حملات را خودکار می‌کند، که تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا نقاط تزریق خاص را با دقت بالا شناسایی کنند و حداکثر تأثیر را داشته باشند.
- **انطباق‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در منطق برنامه‌ها، اقدامات دفاعی یا شرایط محیطی سازگار شوند.
- **پنهان‌کاری:**
  - با تقلید از ترافیک قانونی یا رمزنگاری payloadها تا زمان اجرا، هوش مصنوعی اطمینان می‌دهد که پرس‌وجوهای مخرب شناسایی نشوند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین برنامه وب را به‌طور همزمان اسکن و مورد سوءاستفاده قرار دهند، که پتانسیل آسیب را افزایش می‌دهد.

۲-۲۵-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های برچسب‌دار از آسیب‌پذیری‌های شناخته‌شده تزریق SQL و payloadهای موفق آموزش می‌دهد تا الگوها را یاد بگیرد و موارد جدید را شناسایی کند.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در رفتار برنامه‌ها یا پاسخ‌های پرس‌وجو را شناسایی می‌کند تا آسیب‌پذیری‌های پنهان را آشکار کند.
  - یادگیری تقویتی: تولید payloadها را با پاداش دادن به سوءاستفاده موفق و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، الگوهای پیچیده در

ورودی‌ها، خروجی‌ها و تعاملات برنامه‌ها را تحلیل می‌کنند تا استراتژی‌های حمله را بهبود بخشند.

- شبکه‌های مولد تخاصمی: payloadهای مصنوعی اما واقع‌گرایانه ایجاد می‌کنند که برای دور زدن ابزارهای تحلیل ایستا یا دفاع‌های مبتنی بر امضا طراحی شده‌اند.

- **تکنیک‌های فازی:**
  - هوش مصنوعی را با روش‌های فازی ترکیب می‌کند تا برنامه‌ها را به‌طور سیستماتیک برای رفتارهای غیرمنتظره یا آسیب‌پذیری‌ها تست کند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی مانند پیام‌های خطا، پاسخ‌های API یا مستندات را تحلیل می‌کند تا زمینه‌های بیشتری درباره نقاط تزریق بالقوه استنباط کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا انحرافات را شناسایی کند یا موجودیت‌های قانونی را تقلید کند.

۳-۲۵-۲ دارایی‌های هدف

- **برنامه‌های وب:**
  - برنامه‌هایی با ورودی‌های اعتبارسنجی‌نشده، مانند نوارهای جستجو، فرم‌های ورود به سیستم یا بخش‌های نظرات، که در آن‌ها می‌توان پرس‌وجوهای SQL تزریق کرد.
- **پایگاه‌داده‌ها:**
  - پایگاه‌داده‌های بک‌اند که اطلاعات حساس، از جمله داده‌های شخصی، سوابق مالی یا مالکیت فکری را ذخیره می‌کنند.
- **سیستم‌های سازمانی:**
  - اینترانت‌های شرکتی، سیستم‌های مدیریت ارتباط با مشتری یا پورتال‌های کارمندی که در برابر تزریق SQL آسیب‌پذیر هستند.
- **ادغام‌های شخص ثالث:**
  - افزونه‌ها، ویجت‌ها یا APIهایی که در سیستم‌های بزرگتر ادغام شده‌اند و نقاط ورودی برای حملات تزریق SQL فراهم می‌کنند.
- **زیرساخت‌های حیاتی:**
  - رابط‌های وب‌محور که شبکه‌های برق، سیستم‌های بهداشتی یا شبکه‌های حمل‌ونقل را

کنترل می‌کنند و در آن‌ها تزریق SQL می‌تواند آسیب‌های قابل توجهی ایجاد کند.

۲-۲۵-۴ آسیب‌پذیری‌ها

- **اعتبارسنجی ضعیف ورودی‌ها:**
  - فقدان اعتبارسنجی یا پاک‌سازی قوی ورودی‌های کاربر، اجازه می‌دهد پرس‌وجوهای SQL مخرب بدون بررسی عبور کنند.
- **ساختار نامناسب پرس‌وجوها:**
  - استفاده از رشته‌های الحاق‌شده یا پرس‌وجوهای دینامیک بدون ورودی‌های پارامتری‌شده، برنامه‌ها را در برابر خطرات تزریق آسیب‌پذیر می‌کند.
- **نرم‌افزارهای قدیمی:**
  - استفاده از کتابخانه‌ها، فریم‌ورک‌ها یا سیستم‌های عامل قدیمی که حاوی آسیب‌پذیری‌های شناخته‌شده قابل سوءاستفاده از طریق تزریق SQL هستند.
- **تست امنیتی ناکافی:**
  - تست ناکافی برای تزریق SQL در طول توسعه، برنامه‌ها را در معرض سوءاستفاده قرار می‌دهد.
- **خطای انسانی:**
  - توسعه‌دهندگان ممکن است از طریق روش‌های ضعیف کدنویسی یا پیاده‌سازی نادرست اقدامات امنیتی، آسیب‌پذیری‌ها را ایجاد کنند.

۲-۲۵-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای نقشه‌برداری از برنامه وب هدف استفاده می‌کند و نقاط تزریق بالقوه مانند فیلدهای فرم، URLها یا کوکی‌ها را شناسایی می‌کند.
- **اسکن آسیب‌پذیری:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل برنامه از نظر ضعف‌هایی مانند اعتبارسنجی نادرست ورودی‌ها یا پاک‌سازی ضعیف مستقر می‌کند.
- **تولید payload:**
  - مهاجم از یادگیری عمیق یا شبکه‌های مولد تخاصمی برای ایجاد payloadهای سفارشی که متناسب با آسیب‌پذیری‌های شناسایی‌شده هستند، استفاده می‌کند.

- **اجرای تزریق:**
  - مهاجم payload طراحی‌شده را از طریق نقطه تزریق انتخاب‌شده به برنامه هدف تحویل می‌دهد و اطمینان حاصل می‌کند که با ترافیک عادی ترکیب می‌شود.
- **استخراج داده:**
  - مهاجم پرس‌وجوی SQL تزریق‌شده را اجرا می‌کند تا اطلاعات حساس را استخراج کند، رکوردها را دستکاری کرده یا دسترسی‌ها را افزایش دهد.
- **پایداری:**
  - مهاجم درهای پشتی ایجاد می‌کند یا نقاط دسترسی را حفظ کرده تا امکان سوءاستفاده مکرر در طول زمان فراهم شود.

۶-۲۵-۲ نشانه‌های احتمال نفوذ

- **پرس‌وجوهای غیرمعمول پایگاه‌داده:**
  - ناهنجاری‌ها در لاگ‌های پایگاه‌داده، مانند پرس‌وجوها یا دستورات غیرمنتظره که با عملیات قانونی مرتبط نیستند.
- **افزایش پاسخ‌های خطا:**
  - افزایش ناگهانی در پاسخ‌های خطا از برنامه، که ممکن است نشان‌دهنده تلاش‌های ناموفق تزریق باشد.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربران ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های برنامه یا سرور مربوط به احراز هویت ناموفق، فعالیت غیرعادی یا خطاهای غیرمنتظره.
- **الگوهای دسترسی غیرمنتظره به داده:**
  - انحرافات در الگوهای دسترسی به داده، مانند دسترسی غیرمجاز به جداول یا رکوردهای حساس.

۲-۲۵-۷ تأثیرات

- **نشت داده‌ها:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، سوابق مالی یا مالکیت فکری، که منجر به ضررهای مالی یا آسیب‌های اعتباری می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به‌دلیل پایگاه‌داده‌های به‌خطرافتاده یا رکوردهای دستکاری‌شده.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، کلاهبرداری یا اختلال در خدمات که باعث آسیب‌های مالی می‌شوند.
- **آسیب به شهرت:**
  - از دست دادن اعتماد مشتریان و ارزش برند پس از یک نفوذ یا سوءاستفاده از داده‌های سرقت‌شده.
- **جریمه‌های نظارتی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

۲-۲۵-۸ راه‌های مقابله

- **اعتبارسنجی ورودی‌ها:**
  - قوانین سخت‌گیرانه اعتبارسنجی ورودی‌ها را پیاده‌سازی کنید تا اطمینان حاصل شود که تمام ورودی‌های کاربر با فرمت‌ها و مقادیر مورد انتظار مطابقت دارند.
- **دستورات آماده‌شده:**
  - از پرس‌وجوهای پارامتری‌شده یا دستورات آماده‌شده استفاده کنید تا از دستکاری مستقیم پرس‌وجوهای SQL جلوگیری کنید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، کتابخانه‌ها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده را برطرف کنید.
- **فایروال برنامه وب:**
  - راه‌حل‌های پیشرفته فایروال برنامه وب را مستقر کنید که قادر به شناسایی و مسدود کردن تلاش‌های تزریق SQL در زمان واقعی هستند.

- **تست امنیتی:**
  - تست نفوذ، ارزیابی آسیب‌پذیری و بررسی کد را به‌طور منظم انجام دهید تا ضعف‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **نظارت رفتاری:**
  - از تحلیل رفتاری مبتنی بر هوش مصنوعی استفاده کنید تا الگوهای فعالیت غیرعادی که نشان‌دهنده تلاش‌های تزریق SQL هستند، شناسایی شوند.
- **سرکوب خطاها:**
  - پیام‌های خطای دقیق را سرکوب کنید که ممکن است به مهاجمان بینش‌های ارزشمندی درباره ساختار پایگاه‌داده ارائه دهند.
- **اصل حداقل دسترسی:**
  - اصل حداقل دسترسی را برای حساب‌های پایگاه‌داده اعمال کنید تا تأثیر حملات موفق را محدود کنید.
- **آموزش و آگاهی:**
  - توسعه‌دهندگان و اپراتورها را در مورد روش‌های کدنویسی ایمن و اهمیت اعتبارسنجی و پاک‌سازی ورودی‌ها آموزش دهید.
- **رمزگذاری پایگاه‌داده:**
  - داده‌های حساس ذخیره‌شده در پایگاه‌داده‌ها را رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده را به حداقل برسانید.

۲-۲۶ جعل صوتی و تصویری مبتنی بر هوش مصنوعی (جعل عمیق[1])

جعل صوتی و تصویری مبتنی بر هوش مصنوعی، که معمولاً به عنوان جعل عمیق شناخته می‌شود، شامل استفاده از هوش مصنوعی برای ایجاد محتوای رسانه‌ای بسیار واقع‌گرایانه اما مصنوعی است که ظاهر یا صدای افراد واقعی را تقلید می‌کند. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند محتوای جعل عمیق را با دقت بی‌سابقه‌ای ایجاد کنند و افراد را فریب دهند، افکار عمومی را دستکاری کنند یا مرتکب کلاهبرداری شوند. این نوع حمله به‌خصوص خطرناک است زیرا اعتماد به رسانه‌های دیجیتال را تضعیف می‌کند و تشخیص واقعیت از دروغ را دشوار می‌سازد.

جعل عمیق‌ها به‌طور مخرب در زمینه‌هایی مانند کمپین‌های اطلاعات نادرست، اخاذی، سرقت هویت یا حملات مهندسی اجتماعی استفاده می‌شوند. خودکارسازی و واقع‌گرایی ارائه‌شده توسط

1. Deepfakes

هوش مصنوعی این تهدیدات را به‌طور فزاینده‌ای در دسترس و تشخیص آن‌ها را دشوارتر می‌کند.

۲-۲۶-۱ ویژگی‌های کلیدی

- **واقع‌گرایی بالا:**
  - هوش مصنوعی جعل عمیق‌هایی ایجاد می‌کند که تقریباً از محتوای صوتی یا تصویری واقعی غیرقابل تشخیص هستند و اعتبار و تأثیر آن‌ها را افزایش می‌دهند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند ایجاد، بهبود و توزیع جعل عمیق‌ها را خودکار می‌کنند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهند.
- **انعطاف‌پذیری:**
  - جعل عمیق‌های مبتنی بر هوش مصنوعی به‌طور پویا تکامل می‌یابند و با تغییرات در تکنیک‌های تشخیص، فرمت‌های رسانه‌ای یا شرایط محیطی سازگار می‌شوند.
- **هدف‌گیری دقیق:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند جعل عمیق‌ها را برای افراد، صنایع یا زمینه‌های خاص سفارشی‌سازی کنند تا حداکثر تأثیر را داشته باشند.
- **مقیاس‌پذیری:**
  - تولید خودکار به مهاجمان امکان می‌دهد حجم زیادی از محتوای جعل عمیق را به‌طور کارآمد ایجاد کنند و چندین سیستم یا مخاطب را به‌طور همزمان هدف قرار دهند.

۲-۲۶-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری عمیق:**
  - شبکه‌های مولد تخاصمی: با قرار دادن دو شبکه عصبی در مقابل یکدیگر، محتوای صوتی و تصویری مصنوعی اما واقع‌گرایانه ایجاد می‌کنند—یکی محتوای جعلی ایجاد می‌کند و دیگری آن را تشخیص می‌دهد.
  - رمزگذار خودکار (Autoencoders): داده‌ها را فشرده و بازسازی می‌کنند تا جعل عمیق‌های باکیفیت ایجاد کنند و ویژگی‌های کلیدی مانند حالات چهره یا تن صدا را حفظ کنند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند رونوشت‌ها یا اسکریپت‌ها، را تحلیل می‌کند تا صداگذاری‌ها یا زیرنویس‌های واقع‌گرایانه برای ویدیوهای جعل عمیق ایجاد کند.

- **بینایی کامپیوتری:**
  - از شبکه‌های عصبی پیچشی برای تحلیل و دستکاری محتوای بصری، مانند حالات چهره، حرکات یا پس‌زمینه‌ها، استفاده می‌کند.
- **یادگیری تقویتی:**
  - تولید جعل عمیق را با پاداش‌دهی به فریب موفق و جریمه‌کردن شکست‌ها بهینه می‌کند و امکان بهبود مداوم را فراهم می‌کند.
- **افزایش داده:**
  - مجموعه‌داده‌های آموزشی را با تغییرات داده‌های واقعی گسترش می‌دهد تا استحکام و انطباق‌پذیری مدل را بهبود بخشد.

۳-۲۶-۲ دارایی‌های هدف

- **افراد:**
  - افراد مشهور، سیاستمداران، مدیران اجرایی یا سایر شخصیت‌های عمومی که اعتبار یا حرفه آن‌ها ممکن است توسط جعل عمیق‌ها آسیب ببینند.
- **سازمان‌ها:**
  - شرکت‌هایی که مالکیت فکری، اسرار تجاری یا داده‌های مشتریان ارزشمند را ذخیره می‌کنند و در برابر حملات مبتنی بر جعل عمیق آسیب‌پذیر هستند.
- **سازمان‌های دولتی:**
  - اهداف باارزش درگیر در امنیت ملی، دیپلماسی یا حکومت که در آن‌ها جعل عمیق‌ها می‌توانند آسیب‌های قابل توجهی ایجاد کنند.
- **سیستم‌های مالی:**
  - پورتال‌های بانکی، دروازه‌های پرداخت یا صرافی‌های ارز دیجیتال که در آن‌ها جعل عمیق‌ها می‌توانند برای کلاهبرداری یا دسترسی غیرمجاز استفاده شوند.
- **پلتفرم‌های رسانه‌ای:**
  - شبکه‌های اجتماعی، رسانه‌های خبری یا پلتفرم‌های سرگرمی که در آن‌ها جعل عمیق‌ها می‌توانند اطلاعات نادرست را در مقیاس گسترده منتشر کنند.

۲-۲۶-۴ آسیب‌پذیری‌ها

- **مکانیسم‌های احراز هویت ضعیف:**
  - سیستم‌های بیومتریک که به تشخیص چهره یا تأیید صدا متکی هستند، در برابر جعل جعل عمیق آسیب‌پذیر هستند.
- **روانشناسی انسانی:**
  - افراد بیشتر احتمال دارد به چیزی که می‌بینند یا می‌شنوند اعتماد کنند، که آن‌ها را در برابر فریب مبتنی بر جعل عمیق آسیب‌پذیر می‌کند.
- **مدیریت محتوای ناکافی:**
  - عدم وجود ابزارها یا سیاست‌های قوی مدیریت در پلتفرم‌ها باعث می‌شود جعل عمیق‌ها بدون کنترل منتشر شوند.
- **اتکای بیش از حد به خودکارسازی:**
  - سیستم‌های تشخیص سنتی ممکن است نتوانند ماهیت ظریف جعل عمیق‌های تولیدشده توسط هوش مصنوعی را تشخیص دهند.
- **داده‌های عمومی:**
  - پایگاه‌داده‌های نشت‌یافته، پروفایل‌های شبکه‌های اجتماعی یا سایر منابع عمومی مواد ارزشمندی برای ایجاد جعل عمیق‌های متقاعدکننده فراهم می‌کنند.

۲-۲۶-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری داده‌های عمومی، مانند عکس‌ها، ویدیوها یا ضبط‌های صوتی فرد هدف استفاده می‌کند.
- **آموزش مدل:**
  - مدل‌های یادگیری ماشین، به‌ویژه شبکه‌های مولد تخاصمی، را روی داده‌های جمع‌آوری‌شده آموزش می‌دهد تا الگوها را یاد بگیرد و نسخه‌های مصنوعی ایجاد کند.
- **ایجاد جعل عمیق:**
  - محتوای صوتی یا تصویری جعل عمیق را ایجاد می‌کند که برای اهداف خاص، مانند انتشار اطلاعات نادرست، کلاهبرداری یا جعل هویت افراد مورد اعتماد، سفارشی‌سازی شده است.

- **استقرار:**
  - محتوای جعل عمیق را در چندین پلتفرم توزیع می‌کند و اطمینان می‌یابد که به مخاطبان مورد نظر می‌رسد.
- **بهره‌برداری:**
  - از جعل عمیق برای اهداف مخرب، مانند تأثیرگذاری بر افکار عمومی، اخاذی از افراد یا دسترسی غیرمجاز به سیستم‌ها، استفاده می‌کند.
- **پایداری:**
  - جعل عمیق را بر اساس بازخورد به‌طور مداوم بهبود می‌بخشد و تاکتیک‌ها را تشدید یا محتوا را تغییر می‌دهد تا تأثیر بلندمدت را حفظ کند.

۲-۲۶-۶ نشانه‌های احتمال نفوذ

- **آثار ظریف:**
  - ناسازگاری‌ها در حرکات چهره، الگوهای پلک زدن، نور یا کیفیت صدا که ممکن است نشان‌دهنده وجود جعل عمیق باشد.
- **الگوهای ارسال غیرعادی:**
  - ناهنجاری‌ها در فرکانس ارسال، لحن یا سبک که از رفتار عادی کاربر منحرف می‌شوند.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با عادات یا ترجیحات شناخته‌شده هدف، مانند اظهارات یا اقدامات غیرمنتظره.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به تلاش‌های دسترسی غیرمجاز با استفاده از اعتبارنامه‌های جعل عمیق.
- **نتایج غیرمنتظره:**
  - انحراف در تصمیم‌گیری یا واکنش‌های عمومی که ممکن است تحت تأثیر محتوای جعل عمیق باشد.

۲-۲۶-۷ تأثیرات

- **تفرقه‌افکنی اجتماعی:**
  - جعل عمیق‌ها اطلاعات نادرست را منتشر می‌کنند، افکار عمومی را قطبی می‌کنند و اعتماد

به نهادها یا رسانه‌ها را تضعیف می‌کنند.

- **ضررهای مالی:**
  - فعالیت‌های کلاهبرداری که توسط جعل عمیق امکان‌پذیر می‌شوند، مانند درخواست‌های جعلی انتقال وجه یا سرقت هویت، منجر به آسیب مالی می‌شوند.
- **آسیب به اعتبار:**
  - سازمان‌ها یا افرادی که از حملات مبتنی بر جعل عمیق رنج می‌برند ممکن است اعتبار یا اعتماد را از دست بدهند.
- **اختلال در عملیات:**
  - جعل عمیق‌ها می‌توانند عملیات تجاری، فرآیندهای قانونی یا روابط دیپلماتیک را با معرفی شواهد گمراه‌کننده مختل کنند.
- **آسیب روانی:**
  - قربانیان حملات مبتنی بر جعل عمیق ممکن است استرس، اضطراب یا از دست دادن اعتبار را به دلیل محتوای جعلی تجربه کنند.

۸-۲۶-۲ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به شناسایی آثار ظریف یا ناسازگاری‌ها در رسانه‌های جعل عمیق از طریق تشخیص الگوهای پیشرفته هستند.
- **واترمارک دیجیتال:**
  - شناسه‌های منحصر به فرد یا واترمارک‌ها را در رسانه‌های معتبر جاسازی کنید تا اصالت آن‌ها تأیید شود و دستکاری تشخیص داده شود.
- **احراز هویت محتوا:**
  - مکانیسم‌های تأیید مبتنی بر بلاک‌چین یا رمزنگاری را پیاده‌سازی کنید تا اصالت محتوای رسانه‌ای تضمین شود.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های مکرر از محتوای رسانه‌ای، به‌ویژه در زمینه‌های حساس مانند روزنامه‌نگاری یا اجرای قانون، انجام دهید تا جعل عمیق‌های بالقوه را شناسایی کنید.

- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت که نشان‌دهنده استفاده یا توزیع جعل عمیق است، استفاده کنید.
- **آموزش و آگاهی:**
  - کارکنان، کاربران و عموم مردم را آموزش دهید تا محتوای رسانه‌ای را به‌طور انتقادی ارزیابی کنند و علائم جعل عمیق را تشخیص دهند.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور جعل عمیق و تهدیدات انطباق‌پذیر مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حملات جعل عمیق را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.
- **تلاش‌های مشارکتی:**
  - همکاری بین دولت‌ها، شرکت‌های فناوری و محققان را تقویت کنید تا استانداردها و راه‌حل‌های مشترک برای مقابله با جعل عمیق‌ها توسعه یابد.
- **دفاع پیشگیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های در حال تکامل جعل عمیق استفاده کنید.

۲-۲۷ بازپخش

حمله بازپخش مبتنی بر هوش مصنوعی[1] تکنیک‌های سنتی حمله بازپخش را با قابلیت‌های پیشرفته هوش مصنوعی ترکیب می‌کند تا اثربخشی، پنهان‌کاری و انعطاف‌پذیری آن را افزایش دهد. برخلاف حملات بازپخش سنتی که تنها بر ضبط و ارسال مجدد بسته‌های داده تکیه می‌کنند، نسخه‌های مبتنی بر هوش مصنوعی از الگوریتم‌های یادگیری ماشین برای تحلیل، پیش‌بینی و دستکاری الگوهای ارتباطی در زمان واقعی استفاده می‌کنند. این موضوع باعث می‌شود این حملات بسیار خطرناک‌تر باشند، زیرا می‌توانند با تقلید از رفتار کاربران قانونی یا سوءاستفاده از آسیب‌پذیری‌های ظریف، بسیاری از اقدامات امنیتی سنتی را دور بزنند. این نوع حملات به‌ویژه در اکوسیستم‌های دیجیتال مدرن که اتوماسیون، دستگاه‌های اینترنت اشیا (IoT)، خدمات ابری و زیرساخت‌های حیاتی به شدت

1. AI-Driven Replay Attack

به پروتکل‌های ارتباطی امن وابسته‌اند، اهمیت بیشتری پیدا می‌کنند. ادغام هوش مصنوعی در این حملات نشان‌دهنده تحولی بزرگ در تهدیدات سایبری است که نیازمند مکانیزم‌های دفاعی به‌همان اندازه پیشرفته است.

۱-۲۷-۲ ویژگی‌های کلیدی

- **خودکارسازی:**
  - سیستم‌های مبتنی بر هوش مصنوعی نیاز به دخالت دستی را از بین می‌برند و به مهاجمان اجازه می‌دهند عملیات خود را به طور همزمان روی چندین هدف اجرا کنند.
  - ابزارهای خودکار می‌توانند به طور مداوم شبکه‌ها را برای یافتن فرصت‌های حمله اسکن کنند، خطاهای انسانی را کاهش داده و کارایی را افزایش دهند.
- **بهینه‌سازی:**
  - مدل‌های یادگیری ماشین به حملات اجازه می‌دهند بر اساس بازخورد تلاش‌های قبلی تکامل یابند. به عنوان مثال، یادگیری تقویتی می‌تواند زمان‌بندی و محتوای بازپخش‌ها را بهینه کند تا از تشخیص جلوگیری شود.
  - هوش مصنوعی می‌تواند با تغییرات در پیکربندی شبکه، روش‌های رمزنگاری یا پروتکل‌های احراز هویت سازگار شود و اطمینان حاصل کند که حمله حتی با بهبود دفاع‌ها همچنان مؤثر باقی می‌ماند.
- **پنهان‌کاری:**
  - با تحلیل داده‌های تاریخی، هوش مصنوعی می‌تواند رفتار عادی کاربران را تقلید کند و تشخیص فعالیت‌های مخرب از فعالیت‌های قانونی را برای سیستم‌های تشخیص نفوذ (IDS) دشوارتر کند.
  - تکنیک‌هایی مانند شبکه‌های مولد تخاصمی می‌توانند الگوهای ترافیکی مصنوعی اما واقعی‌نما ایجاد کنند که از تشخیص مبتنی بر امضا جلوگیری می‌کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تعداد زیادی از سیستم‌ها یا دستگاه‌ها را به طور همزمان هدف قرار دهند و تأثیر بالقوه حمله را افزایش دهند.
  - مقیاس‌پذیری به‌ویژه در محیط‌هایی مانند شبکه‌های اینترنت اشیا نگران‌کننده است، جایی که هزاران دستگاه به هم متصل ممکن است آسیب‌پذیر باشند.

- **تصمیم‌گیری در زمان واقعی:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند به‌صورت آنی به شرایط متغیر، مانند وصله‌های امنیتی جدید یا قوانین به‌روزرسانی‌شده فایروال، پاسخ دهند و یکپارچگی حمله را بدون وقفه حفظ کنند.

۲-۲۷-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها با استفاده از مجموعه‌داده‌های برچسب‌دار آموزش می‌بینند تا الگوهای ارتباطات قانونی، مانند دنباله‌های خاصی از فراخوانی‌های API یا توکن‌های احراز هویت، را تشخیص دهند.
  - یادگیری بدون نظارت: ناهنجاری‌ها در ترافیک شبکه، مانند افزایش غیرعادی در حجم درخواست‌ها یا انحراف از رفتارهای مورد انتظار، را شناسایی می‌کند که ممکن است نشان‌دهنده ضعف‌های قابل سوءاستفاده باشد.
  - یادگیری تقویتی: با پاداش دادن به نتایج موفق و جریمه کردن شکست‌ها، استراتژی‌های حمله را بهینه می‌کند و به سیستم اجازه می‌دهد رویکرد خود را در طول زمان بهبود بخشد.
- **یادگیری عمیق:**
  - شبکه‌های عصبی مجموعه‌داده‌های پیچیده، مانند ساختار بسته‌های رمزنگاری‌شده یا داده‌های بیومتریک مبتنی بر تصویر، را پردازش می‌کنند تا الگوهای پنهانی که ممکن است از طریق تحلیل سنتی آشکار نشوند، را کشف کنند.
  - مدل‌های یادگیری عمیق همچنین می‌توانند انواع مختلف ترافیک شبکه را طبقه‌بندی کنند و به مهاجمان کمک کنند روی اهداف باارزش تمرکز کنند.
- **شبکه‌های مولد تخاصمی:**
  - جریان‌های داده مصنوعی اما واقعی‌نما، مانند درخواست‌های ورود جعلی یا قرائت‌های سنسور تقلبی، ایجاد می‌کنند تا سیستم‌های تشخیص را فریب دهند.
  - شبکه‌های مولد تخاصمی می‌توانند رفتار عادی کاربران را شبیه‌سازی کنند و تشخیص فعالیت‌های مشکوک را برای سیستم‌های تشخیص ناهنجاری دشوار کنند.
- **بینایی کامپیوتر:**
  - در سناریوهایی که شامل مکانیزم‌های احراز هویت بصری، مانند تشخیص چهره یا اسکن کد QR هستند، استفاده می‌شود.

- الگوریتم‌های بینایی کامپیوتر می‌توانند تصاویر یا فیلم‌ها را تحلیل کنند تا اطلاعات مفیدی برای طراحی حملات بازپخش استخراج کنند.

- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند فایل‌های لاگ، ورودی‌های خط فرمان یا تعاملات چت‌بات‌ها، را تحلیل می‌کند تا زمینه را درک کرده و بازپخش‌های متقاعدکننده ایجاد کند.
  - در سناریوهایی که شامل چالش‌های احراز هویت مبتنی بر متن یا رابط‌های گفتاری هستند، مفید است.

۳-۲۷-۲ دارایی‌های هدف

- **پروتکل‌های شبکه:**
  - هر پروتکلی که برای ارتباط بین سیستم‌ها استفاده می‌شود، از جمله HTTP(S)، TCP/IP، CoAP، MQTT و غیره، می‌تواند هدف قرار گیرد.
  - آسیب‌پذیری‌ها اغلب از پیاده‌سازی نادرست یا عدم وجود محافظت‌های مناسب، مانند رمزنگاری یا زمان‌بندی، ناشی می‌شوند.

- **سیستم‌های احراز هویت:**
  - احراز هویت مبتنی بر رمز عبور، سیستم‌های مبتنی بر توکن (مانند OAuth)، احراز هویت مبتنی بر گواهی و سیستم‌های بیومتریک همگی در صورت عدم ایمن‌سازی مناسب، در برابر حملات بازپخش آسیب‌پذیر هستند.
  - ضعف‌ها شامل الگوریتم‌های قابل پیش‌بینی تولید توکن، عدم وجود مقادیر nonce یا مدیریت ناکافی نشست‌ها می‌شوند.

- **تراکنش‌های مالی:**
  - درگاه‌های پرداخت، برنامه‌های بانکی، صرافی‌های ارز دیجیتال و پلتفرم‌های تجارت الکترونیک به دلیل ارزش مالی بالای خود، اهداف اصلی هستند.
  - مهاجمان می‌توانند تراکنش‌های قانونی را بازپخش کنند تا وجوه را سرقت کرده یا عملیات تجاری را مختل کنند.

- **دستگاه‌های اینترنت اشیا (IoT):**
  - لوازم خانگی هوشمند، دستگاه‌های پزشکی، سنسورهای صنعتی و سایر دستگاه‌های اینترنت اشیا اغلب از پروتکل‌های ارتباطی سبک‌وزن استفاده می‌کنند که کارایی را بر امنیت ترجیح می‌دهند.

- این دستگاه‌ها ممکن است فاقد مکانیزم‌های رمزنگاری یا احراز هویت قوی باشند و هدف آسانی برای حملات بازپخش باشند.

- **خدمات ابری:**
  - APIها، خدمات ذخیره‌سازی، برنامه‌های کانتینری و معماری‌های بدون سرور اغلب در معرض اینترنت قرار دارند و برای مهاجمان قابل دسترسی هستند.
  - ارائه‌دهندگان ابری باید کنترل‌های دسترسی قوی و نظارت دقیق را برای جلوگیری از دسترسی غیرمجاز از طریق حملات بازپخش تضمین کنند.

- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، سیستم‌های تأمین آب، شبکه‌های حمل‌ونقل و تأسیسات بهداشتی به ارتباطات امن برای عملکرد صحیح وابسته‌اند.
  - یک حمله بازپخش موفق بر زیرساخت‌های حیاتی می‌تواند عواقب فاجعه‌باری داشته باشد و ایمنی عمومی و امنیت ملی را تحت تأثیر قرار دهد.

۴-۲۷-۲ آسیب‌پذیری‌ها

- **عدم استفاده از زمان‌بندی:**
  - بدون زمان‌بندی، سیستم‌ها نمی‌توانند تأیید کنند که یک پیام اخیراً ارسال شده است یا از یک تعامل قبلی بازپخش می‌شود.
  - مهاجمان می‌توانند با تأخیر در ارسال پیام‌های ضبط‌شده و ارسال مجدد آن‌ها در زمان‌های مناسب، از این ضعف سوءاستفاده کنند.

- **مکانیزم‌های احراز هویت ضعیف:**
  - احراز هویت تک‌عاملی (مانند تنها رمز عبور) یا الگوریتم‌های قابل پیش‌بینی تولید توکن، تقلید از کاربران قانونی را برای مهاجمان آسان‌تر می‌کند.
  - عدم استفاده از احراز هویت چندعاملی خطر دسترسی غیرمجاز را افزایش می‌دهد.

- **رمزنگاری ناکافی:**
  - پروتکل‌های رمزنگاری ضعیف یا نادرست می‌توانند داده‌های حساس را در معرض رهگیری و دستکاری قرار دهند.
  - حتی اگر رمزنگاری استفاده شود، مهاجمان همچنان می‌توانند بسته‌های رمزنگاری‌شده را بازپخش کنند اگر هیچ محافظت اضافی وجود نداشته باشد.

- **ثبت گزارش‌های ناکافی:**
  - روش‌های ضعیف ثبت گزارش‌ها، ردیابی و انتساب حملات را دشوار می‌کند و به مهاجمان زمان بیشتری برای فعالیت بدون تشخیص می‌دهد.
  - گزارش‌ها باید شامل اطلاعات دقیقی درباره هر تراکنش، از جمله زمان‌بندی، آدرس‌های IP و شناسه‌های کاربری باشند.
- **آسیب‌پذیری‌های وصله‌نشده:**
  - آسیب‌پذیری‌های شناخته‌شده در نرم‌افزار یا ریزبرنامه[1] می‌توانند توسط مهاجمان برای دسترسی اولیه قبل از راه‌اندازی حمله بازپخش مورد سوءاستفاده قرار گیرند.
  - وصله‌کردن و به‌روزرسانی منظم برای بستن این شکاف‌ها ضروری است.
- **خطای انسانی:**
  - تنظیمات نادرست فایروال‌ها، پورت‌های باز، اعتبارنامه‌های مشترک و سایر اشتباهات مدیران می‌توانند به طور ناخواسته فرصت‌هایی برای مهاجمان ایجاد کنند.
  - آموزش آگاهی امنیتی برای کاهش این خطرات بسیار مهم است.

۵-۲۷-۲ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای نقشه‌برداری از شبکه هدف، شناسایی میزبان‌های فعال، پورت‌های باز و خدمات در حال اجرا استفاده می‌کند.
  - مهاجم متاداده‌هایی درباره الگوهای ارتباطی، مانند چرخه‌های درخواست/پاسخ معمول و روش‌های احراز هویت رایج، جمع‌آوری می‌کند.
- **جمع‌آوری داده:**
  - مهاجم از ابزارهای شنود تقویت‌شده توسط یادگیری ماشین برای رهگیری و ضبط انتقال‌های داده قانونی استفاده می‌کند.
  - او بر ضبط تعاملات باارزش، مانند درخواست‌های ورود، مجوزهای پرداخت یا دستورات مدیریتی، تمرکز دارد.
- **تحلیل آموزش مدل:**
  - مهاجم از الگوریتم‌های یادگیری ماشین برای تحلیل داده‌های ضبط‌شده استفاده کرده و اجزای قابل استفاده مجدد، مانند توکن‌های نشست، کلیدهای API یا امضاهای رمزنگاری،

1. firmware

را شناسایی می‌کند.

- او سپس مدل‌ها را آموزش داده تا تعاملات آینده را پیش‌بینی کنند یا تغییرات معقولی از داده‌های موجود ایجاد کنند.

- **اجرای بازپخش:**
  - مهاجم ارسال مجدد داده‌های ضبط‌شده را در فواصل زمانی مناسب خودکار کرده و اطمینان حاصل می‌کند که زمان‌بندی با الگوهای استفاده عادی هماهنگ است.
  - مهاجم زمان‌بندی و محتوا را به‌صورت پویا تنظیم کرده تا از ایجاد هشدار یا سوءظن جلوگیری شود.
- **استخراج و پایداری:**
  - مهاجم داده‌ها یا دارایی‌های ارزشمند را از سیستم به خطر افتاده استخراج می‌کند
  - مهاجم به طور مداوم محیط را نظارت کرده و استراتژی حمله را بر اساس نتایج مشاهده‌شده اصلاح می‌کند.
  - مهاجم درهای پشتی ایجاد کرده یا نقاط دسترسی را حفظ می‌کند تا حملات مکرر در طول زمان تسهیل شوند.

۶-۲۷-۲ نشانه‌های احتمال نفوذ

- **الگوهای ترافیکی غیرعادی:**
  - درخواست‌های یکسان مکرر در یک بازه زمانی کوتاه ممکن است نشان‌دهنده تلاش‌های خودکار بازپخش باشد.
  - ناهنجاری‌ها در زمان‌بندی درخواست/پاسخ، مانند تأخیرها یا فواصل نامنظم، می‌تواند نشان‌دهنده فعالیت مخرب باشد.
- **نشست‌های تکراری:**
  - چندین نشست با اعتبارنامه‌ها یا شناسه‌های نشست یکسان ممکن است نشان‌دهنده استفاده مجدد از توکن‌های سرقت‌شده باشد.
- **ورودهای غیرمنتظره:**
  - ورود از آدرس‌های IP ناشناس، مکان‌های جغرافیایی غیرمعمول یا دستگاه‌های خارج از ساعات کاری عادی نیاز به بررسی دارد.
- **استفاده غیرعادی از منابع:**
  - افزایش مصرف CPU، حافظه یا پهنای باند ممکن است نشان‌دهنده فرآیندهای غیرمجاز در

حال اجرا در پس‌زمینه باشد.

- **تلاش‌های ناموفق احراز هویت:**
  - افزایش ناگهانی در تلاش‌های ناموفق ورود به دنبال دسترسی موفق ممکن است نشان‌دهنده حملات بروت فورس یا بازپخش باشد.
- **ناهنجاری‌های ترافیک رمزنگاری‌شده:**
  - تغییرات در کلیدهای رمزنگاری، گواهی‌ها یا مجموعه‌های رمزنگاری که توسط مدیران آغاز نشده‌اند، ممکن است نشان‌دهنده دستکاری باشند.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول کاربران، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول، باید باعث ایجاد هشدار شوند.

۷-۲۷-۲ تأثیرات

- **ضررهای مالی:**
  - انتقال‌های غیرمجاز وجوه، تراکنش‌های تقلبی یا سرقت مالکیت فکری می‌تواند منجر به ضررهای مالی قابل توجهی شود.
  - شرکت‌ها همچنین ممکن است با آسیب به اعتبار و مسئولیت‌های قانونی ناشی از نقض‌ها مواجه شوند.
- **اختلال در عملیات:**
  - سیستم‌های به‌خطر افتاده یا داده‌های خراب‌شده می‌توانند باعث پایین آمدن سامانه شوند و بر بهره‌وری و رضایت مشتری تأثیر بگذارند.
  - در برخی موارد، ممکن است نیاز باشد کل عملیات تا زمان حل مشکل متوقف شود.
- **آسیب به اعتبار:**
  - مشتریان و شرکا ممکن است پس از یک نقض، اعتماد خود را به سازمان از دست بدهند و این بر روابط بلندمدت و درآمد تأثیر بگذارد.
  - پوشش منفی رسانه‌ها می‌تواند آسیب را بیشتر تشدید کند.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، می‌تواند منجر به جریمه‌های سنگین و تحریم‌ها شود.

- سازمان‌ها همچنین ممکن است با دعاوی قضایی از سوی افراد یا کسب‌وکارهای آسیب‌دیده مواجه شوند.

- **خطرات ایمنی:**
  - حملات بر زیرساخت‌های حیاتی، مانند نیروگاه‌ها یا بیمارستان‌ها، می‌تواند تهدید مستقیمی برای جان انسان‌ها ایجاد کند.
  - خرابی‌های ناشی از حملات بازپخش می‌تواند منجر به حوادث، آسیب‌ها یا تلفات جانی شود.

۸-۲۷-۲ راه‌های مقابله

- **پیاده‌سازی احراز هویت قوی:**
  - از احراز هویت چندعاملی استفاده کنید تا نیاز به چندین شکل تأیید، مانند چیزی که می‌دانید (رمز عبور)، چیزی که دارید (توکن) و چیزی که هستید (بیومتریک)، وجود داشته باشد.
  - اسرار و کلیدها را به‌طور منظم تغییر دهید تا فرصت مهاجمان را کاهش دهید.
- **فعال‌سازی زمان‌بندی:**
  - در تمام ارتباطات از زمان‌بندی استفاده کنید تا از بازپخش‌های تأخیری جلوگیری شود.
  - زمان‌بندی‌ها را با زمان فعلی سیستم مقایسه کنید تا تازگی آن‌ها تأیید شود.
- **رمزنگاری تمام ارتباطات:**
  - از پروتکل‌های رمزنگاری قوی مانند TLS 1.3 برای ایمن‌سازی داده‌ها در حال انتقال استفاده کنید.
  - اطمینان حاصل کنید که رمزنگاری به‌درستی پیاده‌سازی شده است، از جمله مدیریت کلید و اعتبارسنجی گواهی.
- **نظارت بر ترافیک شبکه:**
  - سیستم‌های تشخیص نفوذ (IDS) مبتنی بر هوش مصنوعی را برای شناسایی فعالیت‌های مشکوک در زمان واقعی مستقر کنید.
  - از تحلیل‌های رفتاری برای ایجاد خطوط پایه و تشخیص انحرافات استفاده کنید.
- **استفاده از مقادیر نانس:**
  - مقادیر یکتا و یک‌بارمصرف را در هر تراکنش بگنجانید تا از استفاده مجدد جلوگیری شود.
  - نانس‌ها باید غیرقابل پیش‌بینی و به‌طور ایمن ذخیره شوند تا از افشا جلوگیری شود.

- **بررسی‌ها و به‌روزرسانی‌های منظم:**
  - ممیزی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
  - مشکلات شناخته‌شده را به‌سرعت وصله کنید تا نقاط ورود بالقوه برای مهاجمان بسته شوند.
- **تحلیل‌های رفتاری:**
  - از هوش مصنوعی برای تحلیل رفتار کاربران استفاده کنید و فعالیت‌هایی که از هنجارهای تعیین‌شده منحرف می‌شوند، را علامت‌گذاری کنید.
  - تشخیص ناهنجاری را با اطلاعات تهدید ترکیب کنید تا دقت بهبود یابد.
- **محدود کردن نرخ:**
  - تعداد درخواست‌ها در واحد زمان را محدود کنید تا خطر حملات خودکار کاهش یابد.
  - محدودیت نرخ تطبیقی را پیاده‌سازی کنید که بر اساس الگوهای ترافیکی مشاهده‌شده تنظیم شود.
- **آموزش کارکنان:**
  - کارکنان را در مورد بهترین روش‌ها برای ایمن‌سازی سیستم‌ها، تشخیص تلاش‌های فیشینگ و گزارش فعالیت‌های مشکوک آموزش دهید.
  - تمرین‌ها و شبیه‌سازی‌های منظم انجام دهید تا یادگیری تقویت شود و برای حوادث واقعی آماده شوید.

۲-۲۸ تهدیدات پیشرفته پایدار

تهدیدات پیشرفته پایدار (APTs)[1] حملات سایبری پیچیده و بلندمدتی هستند که معمولاً توسط بازیگران حمایت‌شده توسط دولت یا مهاجمان بسیار ماهر انجام می‌شوند. این تهدیدات با هدف نفوذ به شبکه‌ها، باقی‌ماندن به‌صورت پنهان برای مدت‌های طولانی و سرقت داده‌های حساس یا اختلال در عملیات حیاتی طراحی شده‌اند. در APTهای مبتنی بر هوش مصنوعی از هوش مصنوعی برای افزایش دقت، انعطاف‌پذیری و پنهان‌کاری تکنیک‌های سنتی APT استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند شناسایی، بهره‌برداری از آسیب‌پذیری‌ها، فرار از تشخیص و بهبود تاکتیک‌های خود را به‌صورت خودکار و در زمان واقعی انجام دهند.

APTهای مبتنی بر هوش مصنوعی به‌خصوص خطرناک هستند زیرا پایداری و پیچیدگی حملات

1. Advanced Persistent Threats

انسانی را با سرعت و مقیاس‌پذیری خودکارسازی ترکیب می‌کنند. این امر به مهاجمان امکان می‌دهد تا کمپین‌های بزرگ‌مقیاس و چندوجهی را اجرا کنند که حتی از اقدامات امنیتی پیشرفته نیز عبور می‌کنند.

۲-۲۸-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی کل چرخه حیات یک APT، از شناسایی اولیه تا خروج داده‌ها، را خودکار می‌کند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **انعطاف‌پذیری:**
  - APTهای مبتنی بر هوش مصنوعی به‌طور پویا تکامل می‌یابند و با تغییرات در دفاع‌های سیستم، پیکربندی شبکه یا شرایط محیطی سازگار می‌شوند.
- **پنهان‌کاری:**
  - با تقلید از ترافیک قانونی، رمزنگاری payloadها یا بهره‌برداری از نقاط کور در سیستم‌های تشخیص، هوش مصنوعی اطمینان می‌دهد که فعالیت‌ها برای مدت طولانی بدون ایجاد هشدار باقی بمانند.
- **هدف‌گیری دقیق:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند اهداف باارزش، مانند زیرساخت‌های حیاتی، مالکیت فکری یا اسرار دولتی، را شناسایی کنند.
- **حضور بلندمدت:**
  - هوش مصنوعی دسترسی پایدار به سیستم‌های به‌خطرافتاده را با بهبود مداوم تکنیک‌های فرار و حفظ درهای پشتی تسهیل می‌کند.

۲-۲۸-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های حملات APT موفق و ناموفق آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا ناهنجاری‌ها در رفتار سیستم را شناسایی می‌کند تا آسیب‌پذیری‌های بالقوه یا مکانیسم‌های دفاعی را آشکار کند.
  - یادگیری تقویتی: تاکتیک‌های حمله را با پاداش‌دهی به نتایج موفق و جریمه‌کردن شکست‌ها

بهینه می‌کند و امکان بهبود مداوم را فراهم می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، وابستگی‌های زمانی در تعاملات کاربر، ترافیک شبکه یا رفتار برنامه‌ها را تحلیل می‌کنند تا استراتژی‌های حمله را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: الگوهای ترافیک یا رفتارهای مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های فرار را آزمایش و بهبود بخشند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا فعالیت‌های مخرب را به‌طور نامحسوس با فعالیت‌های قانونی ترکیب کند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند ایمیل‌ها، اسناد یا لاگ‌ها، را تحلیل می‌کند تا بینش‌هایی استخراج کند، پیام‌های فیشینگ ایجاد کند یا سبک‌های ارتباطی قانونی را شبیه‌سازی کند.
- **هوش جمعی:**
  - از تکنیک‌های هوش مصنوعی غیرمتمرکز برای هماهنگی اقدامات بین چندین دستگاه یا سیستم به‌خطرافتاده استفاده می‌کند و رفتار جمعی را برای حداکثر تأثیر تقلید می‌کند.

۳-۲۸-۲ دارایی‌های هدف

- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تأسیسات بهداشتی، مؤسسات مالی یا شبکه‌های حمل‌ونقل که دسترسی طولانی‌مدت به آن‌ها می‌تواند آسیب‌های قابل توجهی ایجاد کند.
- **سازمان‌های دولتی:**
  - اهداف باارزش که اطلاعات طبقه‌بندی‌شده یا سیستم‌های امنیت ملی را کنترل می‌کنند.
- **سیستم‌های سازمانی:**
  - شبکه‌های شرکتی که مالکیت فکری، اسرار تجاری یا داده‌های مشتریان را در خود جای داده‌اند.
- **اجزای زنجیره تأمین:**
  - فروشندگان، شرکا یا تأمین‌کنندگان شخص ثالث که حساب‌های به‌خطرافتاده آن‌ها به‌عنوان

نقطه ورود به اکوسیستم‌های بزرگ‌تر عمل می‌کنند.

- **محیط‌های تحقیق و توسعه (R&D):**
  - سازمان‌هایی که تحقیقات پیشرفته انجام می‌دهند و نوآوری‌های سرقت‌شده می‌توانند مزیت‌های استراتژیک برای رقبا فراهم کنند.

۲-۲۸-۴ آسیب‌پذیری‌ها

- **امنیت ضعیف نقاط پایانی:**
  - عدم وجود حفاظت قوی در نقاط پایانی، سیستم‌ها را در برابر نفوذ اولیه و حرکت جانبی آسیب‌پذیر می‌کند.
- **نظارت ناکافی:**
  - روش‌های ضعیف ثبت لاگ یا نظارت ناکافی در زمان واقعی، تشخیص علائم ظریف فعالیت‌های APT را دشوارتر می‌کند.
- **خطای انسانی:**
  - کارکنانی که در دام ایمیل‌های فیشینگ می‌افتند، روی لینک‌های مخرب کلیک می‌کنند یا از رمزهای عبور ضعیف استفاده می‌کنند، به‌طور ناخواسته آسیب‌پذیری‌هایی را ایجاد می‌کنند که توسط APTها قابل بهره‌برداری هستند.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های قدیمی یا دفاع‌های نادرست، سازمان‌ها را در برابر تکنیک‌های انطباق‌پذیر APT آسیب‌پذیر می‌کند.
- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - آسیب‌پذیری‌های کشف‌نشده در نرم‌افزار یا سخت‌افزار، پایگاه‌هایی را برای مهاجمان فراهم می‌کنند تا حضور خود را ایجاد و حفظ کنند.

۲-۲۸-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره هدف، از جمله معماری، پیکربندی‌ها و رفتارهای معمول آن استفاده می‌کند.
- **نفوذ اولیه:**
  - مهاجم از تکنیک‌های فیشینگ هدفمند، بدافزار یا سرقت اعتبارنامه‌های مبتنی بر

هوش مصنوعی استفاده می‌کند تا پایگاهی در محیط هدف ایجاد کند.

- **حرکت جانبی:**
  - مهاجم از یادگیری تقویتی برای نقشه‌برداری از توپولوژی شبکه و شناسایی مسیرهای حرکت در سیستم استفاده می‌کند و در عین حال از تشخیص اجتناب می‌کند.
- **ارتقای دسترسی:**
  - مهاجم از تحلیل رفتاری و اسکن آسیب‌پذیری برای ارتقای دسترسی و دستیابی به دارایی‌های باارزش استفاده می‌کند.
- **خروج داده:**
  - مهاجم از هوش مصنوعی برای بهینه‌سازی روش‌های جمع‌آوری و انتقال داده استفاده می‌کند و اطمینان می‌دهد که اطلاعات حساس به‌طور کارآمد و پنهان استخراج می‌شوند.
- **پایداری:**
  - مهاجم تاکتیک‌ها را بر اساس بازخورد به‌طور مداوم بهبود می‌بخشد و درهای پشتی ایجاد می‌کند یا نقاط دسترسی را برای بهره‌برداری بلندمدت حفظ می‌کند.
- **پاک‌کردن ردپا:**
  - مهاجم لاگ‌ها را تغییر می‌دهد، ردپاها را پاک می‌کند یا از تکنیک‌های دیگر برای جلوگیری از تشخیص و حفظ پنهان‌کاری استفاده می‌کند.

۶-۲۸-۲ نشانه‌های احتمال نفوذ

- **الگوهای غیرعادی دسترسی به داده:**
  - ناهنجاری‌ها در دسترسی به فایل‌ها، پرس‌وجوهای پایگاه داده یا فراخوانی‌های API که با عملیات قانونی مرتبط نیستند.
- **افزایش استفاده از منابع:**
  - مصرف بالاتر CPU، حافظه یا پهنای باند به دلیل شناسایی خودکار یا خروج داده.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار عادی کاربر، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرعادی.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت

ناموفق یا تلاش‌های دسترسی غیرمجاز.

- **ترافیک شبکه غیرمنتظره:**
  - انحراف در ترافیک خروجی، مانند اتصال به آدرس‌های IP یا دامنه‌های ناشناخته، که ممکن است نشان‌دهنده ارتباطات فرمان و کنترل (C2) یا خروج داده باشد.

۲-۲۸-۷ تأثیرات

- **نقض داده:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به ضرر مالی یا آسیب به اعتبار می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل نفوذ طولانی‌مدت و دستکاری.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلال در خدمات که باعث آسیب مالی می‌شوند.
- **آسیب به اعتبار:**
  - قربانیان APTهای مبتنی بر هوش مصنوعی به دلیل داده‌های افشا یا دستکاری‌شده، آسیب‌های اعتباری می‌بینند.
- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا فریم‌ورهای حیاتی که سیستم‌ها را غیرقابل اعتماد یا غیرقابل استفاده می‌کند.
- **خطرات امنیت ملی:**
  - حملات به سازمان‌های دولتی یا زیرساخت‌های حیاتی می‌تواند امنیت ملی یا ایمنی عمومی را به خطر بیندازد.

۲-۲۸-۸ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به شناسایی ناهنجاری‌های ظریف در ترافیک شبکه، رفتار کاربر یا لاگ‌های سیستم هستند که نشان‌دهنده فعالیت‌های APT است.

- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا حرکت جانبی محدود شده و تأثیر نفوذهای موفق کاهش یابد.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای نظارت بر انحرافات در رفتار کاربر یا سیستم استفاده کنید که ممکن است نشان‌دهنده قصد مخرب باشد.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، چارچوب نرم‌افزاری و پیکربندی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده در APTها اصلاح شوند.
- **دفاع چندلایه:**
  - دفاع‌های سنتی مانند فایروال‌ها، سیستم‌های تشخیص/پیشگیری از نفوذ (IDS/IPS) و ابزارهای حفاظت از نقاط پایانی را با راه‌حل‌های مبتنی بر هوش مصنوعی ترکیب کنید.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد شناسایی علائم APTها، مانند تلاش‌های فیشینگ یا رفتار غیرعادی سیستم، آموزش دهید.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور APT و تهدیدات انطباق‌پذیر مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت APTها را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.
- **دفاع پیشگیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های در حال تکامل APT استفاده کنید.
- **رعایت اصول رمزنگاری:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حال انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده به حداقل برسد.

۲-۲۹ توزیع بدافزار

توزیع بدافزار[1] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای خودکارسازی و بهبود فرآیند انتشار نرم‌افزارهای مخرب در شبکه‌ها، سیستم‌ها یا کاربران است. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند حجم عظیمی از داده‌ها را تحلیل کنند تا اهداف آسیب‌پذیر را شناسایی کنند، کانال‌های توزیع را بهینه کنند و از مکانیسم‌های تشخیص فرار کنند. این نوع حمله به‌خصوص خطرناک است زیرا دقت و انعطاف‌پذیری هوش مصنوعی را با پتانسیل مخرب بدافزار ترکیب می‌کند و به مهاجمان امکان می‌دهد payloadهای مخرب را در مقیاسی بی‌سابقه توزیع کنند.

توزیع بدافزار مبتنی بر هوش مصنوعی می‌تواند دفاع‌های سنتی مانند فایروال‌ها، نرم‌افزارهای آنتی‌ویروس یا سیستم‌های تشخیص نفوذ (IDS) را با تقلید از ترافیک قانونی، بهره‌برداری از روانشناسی انسانی و سازگاری پویا با محیط‌های در حال تغییر دور بزنند.

۲-۲۹-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی کل چرخه حیات توزیع بدافزار، از شناسایی هدف تا تحویل payload، را خودکار می‌کند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی به‌مرور زمان تکامل می‌یابند و با تغییرات در پیکربندی سیستم، اقدامات دفاعی یا شرایط محیطی سازگار می‌شوند.
- **هدف‌گیری دقیق:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند افراد، سازمان‌ها یا سیستم‌های خاص با دارایی‌های باارزش یا دفاع‌های ضعیف را شناسایی کنند.
- **پنهان‌کاری:**
  - با رمزنگاری payloadها، تقلید از ترافیک قانونی یا بهره‌برداری از نقاط کور در سیستم‌های تشخیص، هوش مصنوعی اطمینان می‌دهد که بدافزار در طول توزیع کشف نشود.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد بدافزار را به‌طور همزمان در چندین سیستم یا شبکه توزیع کنند و آسیب‌های بالقوه را افزایش دهند.

1. Malware Distribution

۲-۲۹-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های تلاش‌های موفق و ناموفق توزیع بدافزار آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا ناهنجاری‌ها در ترافیک شبکه یا رفتار کاربر را شناسایی می‌کند تا آسیب‌پذیری‌های بالقوه را آشکار کند.
  - یادگیری تقویتی: تاکتیک‌های توزیع را با پاداش‌دهی به نتایج موفق و جریمه‌کردن شکست‌ها بهینه می‌کند و امکان بهبود مداوم را فراهم می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، وابستگی‌های زمانی در تعاملات کاربر یا فعالیت شبکه را تحلیل می‌کنند تا زمان‌بندی و توالی حملات را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: payloadها یا رفتارهای مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های فرار را آزمایش و بهبود بخشند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند ایمیل‌ها، چت‌لاگ‌ها یا پست‌های شبکه‌های اجتماعی، را تحلیل می‌کند تا پیام‌های فیشینگ بسیار شخصی‌سازی‌شده ایجاد کند که احتمال آلودگی را افزایش می‌دهد.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا فعالیت‌های مخرب را به‌طور نامحسوس با محیط ترکیب کند.
- **داده‌کاوی:**
  - از هوش مصنوعی برای استخراج داده‌های عمومی از شبکه‌های اجتماعی، فروم‌ها یا پایگاه‌داده‌های نفوذشده استفاده می‌کند تا پروفایل‌های دقیقی از اهداف بالقوه ایجاد کند.

۲-۲۹-۳ دارایی‌های هدف

- **کاربران فردی:**
  - کارکنان، مدیران یا سایر افراد کلیدی که دستگاه‌های آن‌ها به‌عنوان نقطه ورود برای کمپین‌های بزرگ‌تر عمل می‌کنند.

- **شبکه‌های سازمانی:**
  - اینترانت‌های شرکتی که داده‌های حساس، مالکیت فکری یا اطلاعات مشتریان را ذخیره می‌کنند.
- **خدمات ابری:**
  - برنامه‌ها، ذخیره‌سازی یا منابع محاسباتی میزبانی‌شده در ابر که از طریق اعتبارنامه‌های به‌خطرافتاده قابل دسترسی هستند.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های IoT با منابع محدود که اغلب به‌عنوان بخشی از بات‌نت‌ها یا نقاط ورود برای انتشار بدافزار استفاده می‌شوند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تأسیسات بهداشتی، مؤسسات مالی یا شبکه‌های حمل‌ونقل که در آن‌ها بدافزار می‌تواند آسیب‌های قابل توجهی ایجاد کند.

۴-۲۹-۲ آسیب‌پذیری‌ها

- **امنیت ضعیف نقاط پایانی:**
  - عدم وجود راه‌حل‌های قوی آنتی‌ویروس، فایروال‌ها یا ابزارهای حفاظت از نقاط پایانی، سیستم‌ها را در برابر نفوذ اولیه آسیب‌پذیر می‌کند.
- **خطای انسانی:**
  - کاربرانی که در دام ایمیل‌های فیشینگ می‌افتند، روی لینک‌های مخرب کلیک می‌کنند یا فایل‌های آلوده دانلود می‌کنند، به‌طور ناخواسته سیستم‌های خود را در معرض خطر قرار می‌دهند.
- **سیستم‌های قدیمی:**
  - استفاده از کتابخانه‌ها، چارچوب‌ها یا سیستم‌عامل‌های قدیمی که حاوی آسیب‌پذیری‌های شناخته‌شده قابل بهره‌برداری توسط بدافزار هستند.
- **نظارت ناکافی:**
  - روش‌های ضعیف ثبت لاگ یا عدم نظارت در زمان واقعی، تشخیص و پاسخ به تلاش‌های توزیع بدافزار را دشوار می‌کند.

- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - آسیب‌پذیری‌های کشف‌نشده در نرم‌افزار یا سخت‌افزار، زمینه‌های مناسبی برای بدافزار مبتنی بر هوش مصنوعی فراهم می‌کنند تا پایگاهی ایجاد کنند.

۲-۲۹-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره اهداف بالقوه، از جمله حضور آنلاین، عادات ارتباطی یا فعالیت‌های اخیر آن‌ها استفاده می‌کند.
- **اسکن آسیب‌پذیری:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل محیط هدف و شناسایی نقاط ضعف، مانند آسیب‌پذیری‌های اصلاح‌نشده یا مکانیسم‌های احراز هویت ضعیف، به‌کار می‌گیرد.
- **تولید payload:**
  - مهاجم از یادگیری عمیق یا شبکه‌های مولد تخاصمی برای تولید payloadهای سفارشی‌سازی‌شده متناسب با اهداف یا آسیب‌پذیری‌های خاص استفاده می‌کند.
- **کانال‌های توزیع:**
  - مهاجم استقرار بدافزار را از طریق روش‌های مختلف، مانند ایمیل‌های فیشینگ، کیت‌های اکسپلویت یا وب‌سایت‌های به‌خطرافتاده، خودکار می‌کند.
- **تکنیک‌های فرار:**
  - مهاجم از هوش مصنوعی برای ایجاد الگوهای ترافیک یا رفتارهایی استفاده می‌کند که از تشخیص توسط سیستم‌های امنیتی سنتی و پیشرفته اجتناب کنند.
- **پایداری:**
  - مهاجم درهای پشتی ایجاد می‌کند یا نقاط دسترسی را حفظ می‌کند تا امکان آلودگی‌های مکرر یا کنترل بلندمدت بر سیستم‌های به‌خطرافتاده فراهم شود.

۲-۲۹-۶ نشانه‌های احتمال نفوذ

- **الگوهای ترافیک غیرعادی:**
  - ناهنجاری‌ها در ترافیک ورودی یا خروجی، مانند اتصالات غیرمنتظره به آدرس‌های IP یا دامنه‌های ناشناخته.

- **افزایش استفاده از منابع:**
  - مصرف بالاتر CPU، حافظه یا پهنای باند به دلیل اسکن خودکار یا اجرای بدافزار.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار عادی کاربر، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرعادی.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سرور مربوط به ورودی‌های غیرمنتظره، خطاها یا فعالیت‌های غیرعادی.
- **نتایج غیرمنتظره:**
  - انحراف در خروجی‌های سیستم، پیش‌بینی‌ها یا تصمیم‌گیری‌ها که ممکن است نشان‌دهنده تأثیر بدافزار توزیع‌شده باشد.

۷-۲۹-۲ تأثیرات

- **نقض داده:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به ضرر مالی یا آسیب به اعتبار می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل اعتبارنامه‌های به‌خطرافتاده یا آسیب‌پذیری‌های سوءاستفاده‌شده.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، فعالیت‌های کلاهبرداری یا اختلال در خدمات که باعث آسیب مالی می‌شوند.
- **آسیب به شهرت:**
  - سازمان‌هایی که از آلودگی‌های بدافزاری رنج می‌برند ممکن است اعتماد مشتریان و ارزش برند خود را از دست بدهند.
- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا فریم‌ورهای حیاتی که سیستم‌ها را غیرقابل اعتماد یا غیرقابل استفاده می‌کند.

۸-۲۹-۲ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به شناسایی آثار ظریف یا ناسازگاری‌ها در ترافیک یا محتوای تولیدشده توسط بدافزار هستند.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت که نشان‌دهنده توزیع یا عفونت بدافزار است، استفاده کنید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فریم‌ورها و پیکربندی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده توسط توزیع‌کنندگان بدافزار اصلاح شوند.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش بدافزار محدود شده و تأثیر عفونت‌های موفق کاهش یابد.
- **رعایت اصول رمزنگاری:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حال انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده به حداقل برسد.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد شناسایی علائم توزیع بدافزار، مانند تلاش‌های فیشینگ یا رفتار غیرعادی سیستم، آموزش دهید.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا از انواع نوظهور بدافزار و تهدیدات انطباق‌پذیر مطلع شوید.
- **فایروال برنامه‌های وب:**
  - راه‌حل‌های پیشرفته فایروال برنامه وب را مستقر کنید که قادر به تشخیص و مسدودکردن درخواست‌های حاوی بدافزار در زمان واقعی هستند.
- **دفاع پیشگیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر تکنیک‌های

در حال تکامل توزیع بدافزار استفاده کنید.

- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت آلودگی‌های بدافزاری را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.

۲-۳۰ بدافزارهای خودآموز

بدافزارهای خودآموز مبتنی بر هوش مصنوعی نشان‌دهنده یک الگوی جدید در تهدیدات سایبری هستند. در این نوع بدافزارها، نرم‌افزارهای مخرب از هوش مصنوعی برای یادگیری خودکار، سازگاری و تکامل در پاسخ به محیط اطراف خود استفاده می‌کنند. برخلاف بدافزارهای سنتی که از دستورات از پیش تعریف‌شده پیروی می‌کنند، بدافزارهای خودآموز از تکنیک‌های یادگیری ماشین و یادگیری عمیق برای تحلیل محیط، شناسایی آسیب‌پذیری‌ها و بهینه‌سازی رفتار خود به منظور حداکثر تأثیر استفاده می‌کنند. این نوع بدافزار می‌تواند از دفاع‌ها عبور کند، از تشخیص فرار کند و به شکلی مؤثرتر از قبل گسترش یابد.

توانایی بدافزارهای خودآموز مبتنی بر هوش مصنوعی برای بهبود مستمر، آن‌ها را خطرناک‌تر می‌کند، زیرا می‌توانند حتی پیشرفته‌ترین اقدامات امنیتی را دور بزنند و با محیط‌های در حال تکامل سازگار شوند.

۲-۳۰-۱ ویژگی‌های کلیدی

- **یادگیری خودکار:**
  - از یادگیری تقویتی برای بهبود تاکتیک‌های خود بر اساس بازخورد حاصل از تعامل با سیستم یا شبکه هدف استفاده می‌کند.
- **سازگاری:**
  - به صورت پویا کد، رفتار یا بردارهای حمله خود را تغییر می‌دهد تا از تشخیص توسط نرم‌افزارهای آنتی‌ویروس، سیستم‌های تشخیص نفوذ (IDS) یا ابزارهای تحلیل رفتاری فرار کند.
- **پنهان‌کاری:**
  - فرآیندهای قانونی را تقلید می‌کند یا در فعالیت‌های عادی سیستم پنهان می‌شود تا از ایجاد شک جلوگیری کند.
- **مقیاس‌پذیری:**
  - می‌تواند در چندین دستگاه، شبکه یا سیستم گسترش یابد و در عین حال کنترل و هماهنگی

خود را حفظ کند.

- **دقت هدفمند:**
  - محیط هدف را تحلیل می‌کند تا اقدامات خود را برای اهداف خاص، مانند سرقت داده‌ها، استقرار باج‌افزار یا خراب‌کاری، تنظیم کند.

۲-۳۰-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌هایی از حملات موفق و ناموفق آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا ناهنجاری‌ها در رفتار سیستم را شناسایی می‌کند تا نقاط ضعف بالقوه را کشف کند.
  - یادگیری تقویتی: رفتار خود را با پاداش دادن به نتایج موفق (مانند گسترش بدون تشخیص) و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در کد باینری، ترافیک شبکه یا لاگ‌های سیستم را تحلیل می‌کنند تا تکنیک‌های فرار را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: محموله‌ها یا رفتارهای مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های سازگاری را آزمایش و بهبود بخشند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا بدافزار بتواند به‌راحتی در محیط ادغام شود.
- **سنجش محیط:**
  - از هوش مصنوعی برای تشخیص محیط‌های sandbox، تله‌های امنیتی[1] یا سایر اقدامات دفاعی استفاده می‌کند و رفتار خود را بر این اساس تغییر می‌دهد.
- **جهش کد:**
  - از هوش مصنوعی برای تغییر ساختار کد یا جریان اجرای خود استفاده می‌کند در حالی که عملکرد را حفظ می‌کند و انواع پلی‌مورفیک یا متامورفیک ایجاد می‌کند.

1. honeypots

۳-۳۰-۲ دارایی‌های هدف

- **دستگاه‌های پایانی:**
  - کامپیوترها، گوشی‌های هوشمند، تبلت‌ها و دستگاه‌های اینترنت اشیا (IoT) که بدافزار می‌تواند در آن‌ها نفوذ کرده و گسترش یابد.
- **سیستم‌های شبکه:**
  - شبکه‌های سازمانی، پلتفرم‌های ابری یا سیستم‌های کنترل صنعتی که در برابر گسترش پنهانی آسیب‌پذیر هستند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، مراکز درمانی، مؤسسات مالی یا سیستم‌های حمل‌ونقل که اختلال در آن‌ها می‌تواند آسیب‌های قابل‌توجهی ایجاد کند.
- **اجزای زنجیره تأمین:**
  - کتابخانه‌ها، وابستگی‌ها یا فریم‌ورهای شخص ثالث که در سیستم‌های بزرگتر استفاده می‌شوند و نقطه ورودی برای بدافزارهای خودآموز فراهم می‌کنند.
- **مخازن داده:**
  - پایگاه‌های داده، سرورهای فایل یا سیستم‌های ذخیره‌سازی ابری که اطلاعات ارزشمند مانند مالکیت فکری، اسرار تجاری یا داده‌های مشتری را ذخیره می‌کنند.

۴-۳۰-۲ آسیب‌پذیری‌ها

- **دفاع‌های منسوخ:**
  - راه‌حل‌های آنتی‌ویروس سنتی که بر تشخیص مبتنی بر امضا یا تحلیل ایستا تکیه دارند، به راحتی توسط بدافزارهای سازگار دور زده می‌شوند.
- **نظارت رفتاری ناکافی:**
  - ابزارهای تحلیل رفتاری ناکافی قادر به تشخیص انحرافات ظریف ناشی از بدافزارهای خودآموز نیستند.
- **خطای انسانی:**
  - سیستم‌های نادرست پیکربندی‌شده، اعتبارهای مشترک یا آموزش ناکافی سازمان‌ها را در برابر تهدیدات پیچیده آسیب‌پذیر می‌کند.

- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - آسیب‌پذیری‌های کشف‌نشده در نرم‌افزار یا سخت‌افزار زمینه مناسبی برای استقرار بدافزارهای خودآموز فراهم می‌کنند.
- **عدم وجود اقدامات متقابل:**
  - عدم وجود دفاع‌های مبتنی بر هوش مصنوعی یا اطلاعات تهدیدات پیش‌گیرانه، سیستم‌ها را در برابر تهدیدات در حال تکامل آسیب‌پذیر می‌کند.

۲-۳۰-۵ سناریوی تهدید

- **آلودگی اولیه:**
  - از طریق ایمیل‌های فیشینگ، کیت‌های اکسپلویت یا سایر بردارها، دستگاه هدف را آلوده می‌کند و اغلب به عنوان نرم‌افزار قانونی خود را پنهان می‌کند.
- **تحلیل محیط:**
  - از هوش مصنوعی برای اسکن سیستم آلوده و شناسایی پیکربندی‌ها، دفاع‌ها و آسیب‌پذیری‌های بالقوه استفاده می‌کند.
- **یادگیری و سازگاری:**
  - از یادگیری تقویتی برای بهبود رفتار خود بر اساس بازخورد بلادرنگ استفاده می‌کند تا از تشخیص فرار کند و گسترش خود را بهینه‌سازی کند.
- **گسترش:**
  - به صورت جانبی در شبکه گسترش می‌یابد و دستگاه‌ها یا سیستم‌های بیشتری را هدف قرار می‌دهد در حالی که تشخیص داده نمی‌شود.
- **اجرا:**
  - هدف خود را اجرا می‌کند، مانند سرقت داده‌ها، رمزگذاری فایل‌ها برای باج‌گیری یا ایجاد اختلال در عملیات.
- **پایداری:**
  - با سازگاری مداوم با تغییرات محیط و فرار از اقدامات دفاعی، دسترسی بلندمدت خود را حفظ می‌کند.

۲-۳۰-۶ نشانه‌های احتمال نفوذ

- **تغییرات غیرعادی فایل‌ها:**
  - ناهنجاری‌ها در الگوهای ایجاد، حذف یا تغییر فایل‌ها که با فعالیت‌های قانونی مرتبط نیستند.

- **افزایش استفاده از منابع:**
  - فعالیت بیشتر CPU، حافظه یا I/O دیسک به دلیل اجرای بدافزار در پس‌زمینه.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول سیستم ناسازگار هستند، مانند اتصالات شبکه غیرمنتظره یا افزایش غیرمجاز سطح دسترسی.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به احراز هویت ناموفق، فرآیندهای غیرمعمول یا فعالیت‌های غیرمنتظره.
- **ترافیک شبکه غیرمنتظره:**
  - انحرافات در ترافیک خروجی، مانند اتصالات به آدرس‌های IP یا دامنه‌های ناشناخته.

۷-۳۰-۲ تأثیرات

- **نفوذ به داده‌ها:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری که منجر به زیان‌های مالی یا آسیب به اعتبار می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل سوءاستفاده از اعتبارها یا آسیب‌پذیری‌ها.
- **زیان‌های مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزار یا اختلال در خدمات که باعث آسیب مالی می‌شود.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و اقدامات قانونی می‌شود.
- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا فریم‌ورهای حیاتی که سیستم‌ها را غیرقابل استفاده می‌کند.

۸-۳۰-۲ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - از راه‌حل‌های آنتی‌ویروس مبتنی بر هوش مصنوعی استفاده کنید که بر روی مجموعه‌داده‌های بزرگی از نمونه‌های بدافزار آموزش دیده‌اند تا تهدیدات مبهم را به طور مؤثرتری شناسایی کنند.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی استفاده کنید تا فعالیت‌های غیرعادی که نشان‌دهنده بدافزارهای خودآموز هستند را تشخیص دهید، حتی اگر امضای شناخته‌شده‌ای نداشته باشند.
- **تحلیل پویا:**
  - محیط‌های sandbox را پیاده‌سازی کنید که قادر به اجرا و تحلیل فایل‌های مشکوک در تنظیمات کنترل‌شده هستند تا رفتارهای پنهان را آشکار کنند.
- **به‌روزرسانی منظم:**
  - تمام نرم‌افزارها، فریم‌ورها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده که توسط بدافزارهای خودآموز سوءاستفاده می‌شوند را رفع کنید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچکتر تقسیم کنید تا گسترش بدافزار محدود شود و تأثیر عفونت‌های موفق کاهش یابد.
- **رعایت اصول رمزنگاری:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حالت انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده کاهش یابد.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد علائم عفونت بدافزار، مانند تلاش‌های فیشینگ یا رفتار غیرعادی سیستم آموزش دهید.
- **اطلاعات تهدیدات پیشرفته:**
  - از پلتفرم‌های اطلاعات تهدیدات مبتنی بر هوش مصنوعی استفاده کنید تا در مورد بدافزارهای خودآموز و تهدیدات سازگار به‌روز بمانید.

- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به سرعت حملات بدافزارهای خودآموز را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.
- **دفاع پیش‌گیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر انواع در حال تکامل بدافزار استفاده کنید.

۲-۳۱ کلون‌سازی

حمله کلون‌سازی[1] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای تکثیر یا تقلید رفتار، هویت یا عملکرد یک موجودیت قانونی مانند یک دستگاه، کاربر، برنامه یا مؤلفه سیستم است. این نوع حمله خطرناک است زیرا به مهاجمان امکان می‌دهد سیستم‌ها، کاربران یا شبکه‌ها را فریب دهند تا باور کنند که موجودیت کلون‌شده معتبر است. با استفاده از تکنیک‌های یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند کلون‌هایی با واقع‌گرایی بالا ایجاد کنند که تشخیص آن‌ها از نسخه‌های اصلی دشوار است.

حملات کلون‌سازی معمولاً در حوزه‌هایی مانند دستگاه‌های موبایل (کلون‌کردن سیم‌کارت)، اکوسیستم‌های اینترنت اشیا (کلون‌کردن دستگاه)، هویت‌های دیجیتال (کلون‌کردن حساب کاربری) و حتی سیستم‌های تشخیص صدا یا چهره (کلون‌کردن بیومتریک) مشاهده می‌شوند. هنگامی که این حملات با هوش مصنوعی تقویت می‌شوند، پیچیده‌تر، مقیاس‌پذیرتر و قادر به دور زدن اقدامات امنیتی سنتی می‌شوند.

۱- ۲-۳۱ ویژگی‌های کلیدی

- **تکثیر با دقت بالا:**
  - هوش مصنوعی اطمینان می‌دهد که موجودیت‌های کلون‌شده به‌طور نزدیکی از نظر رفتار، ظاهر یا الگوهای عملیاتی، نسخه اصلی را تقلید می‌کنند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند شناسایی اهداف، جمع‌آوری داده‌ها و تولید کلون‌ها را خودکار می‌کنند و تلاش دستی را کاهش می‌دهند.
- **انعطاف‌پذیری:**
  - کلون‌های مبتنی بر هوش مصنوعی می‌توانند به‌طور پویا با تغییرات محیط، مانند مکانیزم‌های

1. Cloning Attack

احراز هویت به‌روزرسانی‌شده یا خطوط پایه رفتاری، سازگار شوند.

- **پنهان‌کاری:**
  - کلون‌ها به‌گونه‌ای طراحی شده‌اند که به‌راحتی با موجودیت‌های قانونی ترکیب شوند و تشخیص آن‌ها از طریق تکنیک‌های نظارتی معمولی دشوارتر شود.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین موجودیت را به‌طور همزمان کلون کنند و دامنه و تأثیر حمله را افزایش دهند.

۲-۳۱-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های رفتارها یا ویژگی‌های موجودیت‌های قانونی آموزش می‌دهد تا کلون‌های دقیقی ایجاد کند.
  - یادگیری بدون نظارت: الگوهای فعالیت عادی را شناسایی می‌کند تا راهنمایی برای ایجاد کلون‌های متقاعدکننده بدون برچسب‌گذاری قبلی باشد.
  - یادگیری تقویتی: استراتژی‌های کلون‌سازی را با پاداش دادن به جعل‌های موفق و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در داده‌های تصویری، صوتی یا رفتاری را تحلیل می‌کنند تا کلون‌های واقع‌گرایانه تولید کنند.
  - شبکه‌های مولد تخاصمی: نسخه‌های مصنوعی اما واقع‌گرایانه از چهره‌ها، صداها یا اثرانگشت دستگاه‌ها ایجاد می‌کنند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند لاگ‌های چت یا ایمیل‌ها، را تحلیل می‌کند تا سبک‌های ارتباطی و استفاده از زبان را برای اهداف مهندسی اجتماعی تکثیر کند.
- **بینایی کامپیوتر:**
  - در کلون‌سازی بیومتریک برای ایجاد تصاویر یا ویدیوهای جعلی از چهره‌ها، عنبیه‌ها یا اثرانگشت‌ها استفاده می‌شود که می‌توانند سیستم‌های احراز هویت را فریب دهند.

- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار کاربر یا سیستم استفاده می‌کند و امکان ایجاد کلون‌هایی را فراهم می‌کند که مانند نسخه‌های اصلی عمل می‌کنند.

۳-۳۱-۲ دارایی‌های هدف

- **دستگاه‌های موبایل:**
  - سیم‌کارت‌ها، شماره‌های تلفن و شناسه‌های دستگاه که در برابر کلون‌سازی برای دسترسی غیرمجاز یا کلاهبرداری آسیب‌پذیر هستند.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - لوازم خانگی هوشمند، سنسورهای صنعتی یا دستگاه‌های پزشکی که برای احراز هویت به شناسه‌ها یا گواهی‌های منحصر به فرد متکی هستند.
- **هویت‌های دیجیتال:**
  - حساب‌های کاربری، پروفایل‌ها یا اعتبارنامه‌های مورد استفاده در پلتفرم‌های آنلاین، سیستم‌های مالی یا محیط‌های سازمانی.
- **سیستم‌های بیومتریک:**
  - سیستم‌های تشخیص چهره، اسکن اثرانگشت یا تأیید صدا که در برابر جعل‌های تولیدشده توسط هوش مصنوعی آسیب‌پذیر هستند.
- **خدمات ابری:**
  - ماشین‌های مجازی، کانتینرها یا حساب‌های ذخیره‌سازی که می‌توانند برای سرقت داده‌ها یا اختلال در عملیات کلون شوند.
- **زیرساخت‌های حیاتی:**
  - سیستم‌های کنترل، شبکه‌های ارتباطی یا شبکه‌های برق که در آن موجودیت‌های کلون‌شده می‌توانند آسیب‌های قابل توجهی ایجاد کنند.

۴-۳۱-۲ آسیب‌پذیری‌ها

- **مکانیزم‌های احراز هویت ضعیف:**
  - احراز هویت تک‌عاملی یا الگوریتم‌های قابل پیش‌بینی تولید توکن، کلون‌سازی موجودیت‌ها را برای مهاجمان آسان‌تر می‌کند.

- **مکانیزم‌های تأییدی ضعیف :**
  - عدم وجود مکانیزم‌های قوی شناسایی دستگاه، امکان عبور دستگاه‌های کلون‌شده به‌عنوان دستگاه‌های قانونی را فراهم می‌کند.
- **رمزنگاری ناکافی:**

پروتکل‌های رمزنگاری ضعیف یا نادرست، داده‌های حساس مورد نیاز برای کلون‌سازی را در معرض دید قرار می‌دهند.

- **ثبت گزارش‌های ناکافی:**

روش‌های ضعیف ثبت گزارش‌ها، ردیابی و انتساب فعالیت‌های کلون‌سازی را دشوار می‌کند.

- **خطای انسانی:**

تنظیمات نادرست فایروال‌ها، اعتبارنامه‌های مشترک یا سایر اشتباهات مدیران ممکن است به‌طور ناخواسته کلون‌سازی را تسهیل کند.

۵-۳۱-۲ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره موجودیت هدف، از جمله رفتار، ویژگی‌ها و تعاملات آن استفاده می‌کند.
- **پیش‌پردازش:**
  - مهاجم داده‌های مرتبط، مانند متادیتای دستگاه، نمونه‌های بیومتریک یا الگوهای ارتباطی را با استفاده از ابزارهای شنود یا تکنیک‌های مهندسی اجتماعی ضبط می‌کند.
- **تولید کلون:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های جمع‌آوری‌شده و ایجاد یک کلون واقع‌گرایانه از موجودیت هدف مستقر می‌کند.
- **استقرار:**
  - مهاجم موجودیت کلون‌شده را در سیستم یا شبکه هدف معرفی کرده و اطمینان حاصل می‌کند که به‌طور غیرقابل تشخیص از نسخه اصلی رفتار می‌کند.
- **سوءاستفاده:**
  - مهاجم از موجودیت کلون‌شده برای انجام اقدامات غیرمجاز، مانند دسترسی به منابع محدود، سرقت داده‌ها یا اختلال در عملیات استفاده می‌کند.

- **پایداری:**
  - مهاجم حضور موجودیت کلون‌شده را در طول زمان حفظ کرده و با تغییرات محیط سازگار شده تا از تشخیص جلوگیری کند.

۲-۳۱-۶ نشانه‌های احتمال نفوذ

- **موجودیت‌های تکراری:**
  - چندین نمونه از همان دستگاه، کاربر یا حساب که در لاگ‌ها یا پایگاه‌داده‌ها ظاهر می‌شوند.
- **الگوهای فعالیت غیرعادی:**
  - ناهنجاری‌ها در رفتار، مانند ورود همزمان از مکان‌های مختلف یا استفاده غیرمنتظره از منابع.
- **افزایش تلاش‌های احراز هویت:**
  - افزایش ناگهانی در تلاش‌های ناموفق ورود به دنبال دسترسی موفق ممکن است نشان‌دهنده کلون‌سازی باشد.
- **ورودی‌های غیرمنتظره در لاگ‌ها:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مرتبط با دستگاه‌های ناشناخته، دستورات غیرمعمول یا فعالیت غیرعادی.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول کاربران، مانند دسترسی به بخش‌های نامرتبط سیستم یا اجرای دستورات غیرمجاز.

۲-۳۱-۷ تأثیرات

- **سرقت هویت:**
  - دسترسی غیرمجاز به حساب‌های شخصی یا سازمانی، منجر به ضررهای مالی یا آسیب به اعتبار.
- **کلاهبرداری مالی:**
  - ابزارهای پرداخت کلون‌شده، مانند کارت‌های اعتباری یا حساب‌های بانکی، که برای تراکنش‌های کلاهبرداری استفاده می‌شوند.
- **اختلال در عملیات:**
  - دستگاه‌ها یا سیستم‌های کلون‌شده که باعث خرابی، downtime یا از دست دادن خدمات

حیاتی می‌شوند.

- **نقض داده‌ها:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالکیت فکری یا اسرار تجاری.
- **خطرات ایمنی:**
  - تهدید برای جان انسان‌ها در صورت دخالت موجودیت‌های کلون‌شده در زیرساخت‌های حیاتی، دستگاه‌های پزشکی یا سیستم‌های حمل‌ونقل.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

۸-۳۱-۲ راه‌های مقابله

- **احراز هویت چندعاملی:**
  - نیاز به چندین شکل تأیید، مانند چیزی که می‌دانید (رمز عبور)، چیزی که دارید (توکن) و چیزی که هستید (بیومتریک).
- **احراز هویت دستگاه:**
  - مکانیزم‌های قوی شناسایی دستگاه را پیاده‌سازی کنید تا هر موجودیت را به‌طور منحصر به فرد شناسایی و احراز هویت کنید.
- **تحلیل‌های رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای ایجاد خطوط پایه و تشخیص انحرافات نشان‌دهنده کلون‌سازی استفاده کنید.
- **بررسی‌های منظم:**
  - ممیزی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **رمزنگاری و توکن‌سازی:**
  - داده‌های حساس را رمزنگاری کنید و به جای اعتبارنامه‌های ثابت از توکن‌های پویا استفاده کنید تا خطر کلون‌سازی کاهش یابد.
- **کنترل‌های دسترسی:**
  - اصول حداقل دسترسی را اجرا کنید و کنترل‌های دسترسی مبتنی بر نقش (RBAC) را پیاده‌سازی کنید تا تأثیر حملات موفق را محدود کنید.

- **نظارت و ثبت لاگ:**
  - لاگ‌های دقیقی از فعالیت سیستم نگه دارید و با استفاده از ابزارهای نظارتی پیشرفته، رفتارهای مشکوک را نظارت کنید.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد بهترین روش‌ها برای ایمن‌سازی هویت‌های خود و تشخیص علائم کلون‌سازی آموزش دهید.

۲-۳۲ مهندسی معکوس

مهندسی معکوس[1] فرآیند تحلیل یک سیستم، دستگاه یا نرم‌افزار برای درک طراحی، عملکرد یا کارکردهای داخلی آن است. در حمله مهندسی معکوس مبتنی بر هوش مصنوعی، از هوش مصنوعی برای خودکارسازی و بهبود فرآیند تجزیه و بازسازی سیستم‌های هدف استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند باینری‌ها، فرمور یا الگوریتم‌های اختصاصی را با سرعت و دقت بی‌سابقه تحلیل کنند و آسیب‌پذیری‌ها، مالکیت فکری یا اطلاعات حساس را کشف کنند.

این نوع حمله به‌ویژه در صنایعی که فناوری‌های اختصاصی، طرح‌های رمزنگاری یا پروتکل‌های سفارشی جزو دارایی‌های حیاتی هستند، خطرناک است. حملات مهندسی معکوس مبتنی بر هوش مصنوعی می‌توانند با شناسایی الگوها و وابستگی‌هایی که ممکن است توسط انسان‌ها نادیده گرفته شوند، از اقدامات امنیتی سنتی مانند مبهم‌سازی یا رمزنگاری عبور کنند.

۲-۳۲-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند زمان‌بر تجزیه باینری‌ها، دی‌کامپایل کد یا تحلیل فرمور را خودکار می‌کند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **مقیاس‌پذیری:**
  - مدل‌های یادگیری ماشین می‌توانند مجموعه‌های بزرگ داده‌های باینری یا فرمور را تحلیل کنند و به مهاجمان امکان می‌دهند تا چندین سیستم را به‌طور همزمان مهندسی معکوس کنند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در

1. Reverse Engineering attack

نرم‌افزار، سخت‌افزار یا مکانیزم‌های دفاعی مانند مبهم‌سازی کد یا تکنیک‌های ضد دستکاری سازگار شوند.

- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا آسیب‌پذیری‌ها، الگوریتم‌ها یا جریان‌های منطقی خاص را در سیستم‌های پیچیده شناسایی کنند.
- **پنهان‌کاری:**
  - با استفاده از سیگنال‌های ظریف یا روش‌های غیرمستقیم، حملات مهندسی معکوس مبتنی بر هوش مصنوعی می‌توانند حتی در محیط‌های به‌خوبی ایمن‌سازی‌شده نیز کشف نشوند.

۲-۳۲-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌های داده‌ای از باینری‌ها، فرمور یا الگوریتم‌های شناخته‌شده آموزش می‌دهد تا ساختارها و رفتارهای آن‌ها را یاد بگیرد.
  - یادگیری بدون نظارت: الگوها یا خوشه‌ها را در داده‌های بدون برچسب شناسایی می‌کند و به کشف روابط یا عملکردهای پنهان کمک می‌کند.
  - یادگیری تقویتی: فرآیند مهندسی معکوس را با پاداش دادن به کشف‌های موفقیت‌آمیز و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در کد باینری، فرمور یا رفتار پروتکل را تحلیل می‌کنند.
  - شبکه‌های مولد تخاصمی: باینری‌ها یا فرمورهای مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا برای اهداف آموزشی یا آزمایش اقدامات دفاعی استفاده شوند.
- **پردازش زبان طبیعی:**
  - متاداده‌های متنی، نظرات یا مستندات موجود در باینری‌ها یا فرمور را تحلیل می‌کند تا زمینه‌های اضافی درباره سیستم را استنباط کند.
- **اجرای نمادین:**
  - از هوش مصنوعی برای شبیه‌سازی مسیرهای اجرای برنامه استفاده می‌کند و آسیب‌پذیری‌ها یا نقص‌های منطقی بالقوه را بدون نیاز به داده‌های زمان اجرا شناسایی می‌کند.

- **تکنیک‌های :فازی:**
  - هوش مصنوعی را با روش‌های فازی ترکیب می‌کند تا سیستم‌ها را به‌طور سیستماتیک برای رفتارهای غیرمنتظره یا خرابی‌ها آزمایش کند و منطق زیرین را آشکار کند.

۲-۳۲-۳ دارایی‌های هدف

- **نرم‌افزارهای اختصاصی:**
  - برنامه‌ها، کتابخانه‌ها یا چارچوب‌هایی که حاوی مالکیت فکری یا اسرار تجاری ارزشمند هستند.
- **سیستم‌های تعبیه‌شده:**
  - فرمور کنترل‌کننده دستگاه‌های اینترنت اشیا (IoT)، ایمپلنت‌های پزشکی، ماشین‌آلات صنعتی یا سیستم‌های خودرویی.
- **الگوریتم‌های رمزنگاری:**
  - طرح‌های رمزنگاری سفارشی یا پروتکل‌هایی که در نرم‌افزار یا سخت‌افزار پیاده‌سازی شده‌اند.
- **برنامه‌های موبایل:**
  - برنامه‌های موبایلی که داده‌های حساس را ذخیره می‌کنند یا مکانیزم‌های ارتباطی امن را پیاده‌سازی می‌کنند.
- **دستگاه‌های سخت‌افزاری:**
  - تراشه‌ها، سنسورها یا سایر قطعات سخت‌افزاری که دارای فرمور یا میکروکد تعبیه‌شده هستند.

۲-۳۲-۴ آسیب‌پذیری‌ها

- **مبهم‌سازی ضعیف:**
  - مبهم‌سازی کد ضعیف، باینری‌ها یا فرمور را در برابر تحلیل آسیب‌پذیر می‌کند.
- **رمزنگاری ناکافی:**
  - رمزنگاری ضعیف یا نادرست، داده‌ها یا منطق حساس را در معرض مهندسی معکوس قرار می‌دهد.
- **عدم وجود اقدامات ضد دستکاری:**
  - عدم وجود مکانیزم‌هایی مانند بررسی‌های یکپارچگی یا فرآیندهای بوت امن، دسترسی

غیرمجاز به فرمور یا باینری‌ها را تسهیل می‌کند.

- **خطای انسانی:**
  - سیستم‌های نادرست پیکربندی‌شده، اعتبارنامه‌های اشتراکی یا تست ناکافی ممکن است به‌طور ناخواسته آسیب‌پذیری‌هایی را ایجاد کنند که از طریق مهندسی معکوس قابل سوءاستفاده باشند.
- **نشت داده:**
  - باینری‌ها، فرمور یا کد منبع نشت‌یافته، مواد ارزشمندی برای حملات مهندسی معکوس مبتنی بر هوش مصنوعی فراهم می‌کنند.

۵-۳۲-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری باینری‌ها، فرمور یا سایر مواد هدف از منابع عمومی، پایگاه‌های داده نشت‌یافته یا نظارت پنهان استفاده می‌کند.
- **پیش‌پردازش:**
  - مهاجم تکنیک‌های پردازش سیگنال را برای پاک‌سازی و نرمال‌سازی داده‌های جمع‌آوری‌شده اعمال می‌کند، نویز را حذف کرده و ویژگی‌های مرتبط را بهبود می‌بخشد.
- **آموزش مدل:**
  - مهاجم مدل‌های یادگیری ماشین را بر روی داده‌های پیش‌پردازش‌شده آموزش می‌دهد تا الگوها، ساختارها یا آسیب‌پذیری‌ها را در سیستم هدف شناسایی کند.
- **تحلیل:**
  - مهاجم مدل آموزش‌دیده را برای تحلیل سیستم هدف مستقر می‌کند و اطلاعات حساس مانند الگوریتم‌ها، کلیدها یا جریان‌های منطقی را استخراج می‌کند.
- **سوءاستفاده:**
  - مهاجم از اطلاعات استخراج‌شده برای اهداف مخرب، مانند کلون‌کردن دستگاه‌ها، دور زدن اقدامات امنیتی یا سرقت مالکیت فکری استفاده می‌کند.
- **پایداری:**
  - مهاجم فرآیند تحلیل را بر اساس نتایج مشاهده‌شده به‌طور مداوم بهبود می‌بخشد تا موفقیت بلندمدت را تضمین کند.

۲-۳۲-۶ نشانه‌های احتمال نفوذ

- **دسترسی غیرمنتظره به فایل:**
  - ناهنجاری‌ها در الگوهای دسترسی به فایل، مانند تلاش‌های غیرمعمول برای خواندن باینری‌ها یا فرمور.
- **افزایش استفاده از منابع:**
  - مصرف بالاتر CPU یا حافظه به دلیل فعالیت‌های مهندسی معکوس.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول سیستم ناسازگار است، مانند تلاش‌های مکرر برای دسترسی به منابع یا عملکردهای خاص.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به دسترسی غیرمجاز یا فعالیت‌های غیرعادی.
- **ترافیک شبکه غیرمعمول:**
  - انحرافات در ترافیک شبکه، مانند آپلود یا دانلود غیرمنتظره فایل‌های باینری.

۲-۳۲-۷ تأثیرات

- **سرقت مالکیت فکری:**
  - سرقت الگوریتم‌ها، طراحی‌ها یا منطق تجاری اختصاصی که منجر به خسارات مالی یا اختلال رقابتی می‌شود.
- **خطرات امنیتی:**
  - افشای آسیب‌پذیری‌ها یا نقاط ضعف در سیستم‌ها که امکان سوءاستفاده یا حملات بیشتر را فراهم می‌کند.
- **اختلال در عملیات:**
  - سیستم‌های به‌خطر افتاده یا دستگاه‌های کلون‌شده که باعث خرابی، downtime یا از دست‌دادن خدمات حیاتی می‌شوند.
- **خسارات مالی:**
  - دسترسی غیرمجاز به سیستم‌های مالی یا سرقت دارایی‌های دیجیتال.

- **آسیب به اعتبار:**
  - از دست دادن اعتماد مشتری و ارزش برند پس از یک نفوذ یا سرقت مالکیت فکری.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

۸-۳۲-۲ راه‌های مقابله

- **مبهم‌سازی کد:**
  - تکنیک‌های پیشرفته مبهم‌سازی کد را پیاده‌سازی کنید تا تحلیل باینری‌ها یا فرمور دشوارتر شود.
- **رمزنگاری:**
  - داده‌ها، منطق یا کانال‌های ارتباطی حساس را با استفاده از الگوریتم‌های رمزنگاری قوی رمزنگاری کنید.
- **اقدامات ضد دستکاری:**
  - از مکانیزم‌هایی مانند بوت امن، بررسی‌های یکپارچگی یا محیط‌های اجرای مورد اعتماد (TEE)[1] برای جلوگیری از تغییرات غیرمجاز استفاده کنید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرمورها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده را برطرف کنید.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت که نشان‌دهنده تلاش‌های مهندسی معکوس هستند، استفاده کنید.
- **روش‌های توسعه ایمن:**
  - دستورالعمل‌های کدنویسی ایمن را دنبال کنید و تست‌های جامع انجام دهید تا آسیب‌پذیری‌ها را به حداقل برسانید.
- **مدیریت حقوق دیجیتال (DRM):**
  - راه‌حل‌های DRM[2] را پیاده‌سازی کنید تا از نرم‌افزار یا محتوای اختصاصی در برابر دسترسی

1. Transesophageal Echocardiogram (TEE)
2. Digital Rights Management (DRM)

غیرمجاز محافظت کنید.

- **اقدامات امنیتی فیزیکی:**
  - سیستم‌های حیاتی را با موانع فیزیکی، محافظها یا محفظه‌های مقاوم در برابر دستکاری محافظت کنید تا دسترسی به قطعات حساس محدود شود.

۲-۳۳ کانال جانبی

حمله کانال جانبی[1] مبتنی بر هوش مصنوعی از هوش مصنوعی برای بهره‌برداری از کانال‌های جانبی - منابع غیرمستقیم نشت اطلاعات از یک سیستم - استفاده می‌کند تا داده‌های حساس را استخراج یا امنیت را به خطر بیندازد. کانال‌های جانبی شامل اطلاعات زمانی، مصرف انرژی، انتشار الکترومغناطیسی، سیگنال‌های صوتی یا رفتار کش هستند. حملات سنتی کانال جانبی نیاز به تلاش و تخصص زیادی دارند، اما هوش مصنوعی این حملات را با تحلیل حجم زیادی از داده‌های کانال جانبی، شناسایی الگوها و استخراج اسرار با دقت بالا خودکار و بهبود می‌بخشد.

در حمله کانال جانبی مبتنی بر هوش مصنوعی، از هوش مصنوعی برای افزایش دقت و کارایی این تکنیک‌ها با تحلیل مجموعه‌داده‌های بزرگ اطلاعات کانال جانبی، شناسایی الگوها و خودکارسازی استخراج اسرار استفاده می‌شود. این موضوع باعث می‌شود حمله مقیاس‌پذیرتر، انعطاف‌پذیرتر و قادر به غلبه بر اقدامات متقابل سنتی باشد.

۲-۳۳-۱ ویژگی‌های کلیدی

- **تحلیل مبتنی بر داده:**
  - هوش مصنوعی حجم عظیمی از داده‌های کانال جانبی (مانند اندازه‌گیری‌های زمان‌بندی یا ردیابی‌های مصرف برق) را پردازش می‌کند تا همبستگی‌های بین ورودی‌ها و خروجی‌ها را کشف کند.
- **خودکارسازی:**
  - مدل‌های یادگیری ماشین فرآیند شناسایی آسیب‌پذیری‌ها را خودکار می‌کنند و نیاز به تحلیل دستی را کاهش می‌دهند.
- **انعطاف‌پذیری:**
  - هوش مصنوعی می‌تواند با تغییرات در سیستم هدف، مانند الگوریتم‌های به‌روزرسانی‌شده یا اصلاحات سخت‌افزاری، سازگار شود و اطمینان حاصل کند که حمله همچنان مؤثر باقی می‌ماند.

1. Side-Channel Attack

- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا بیت‌ها یا بایت‌های خاصی از داده‌های حساس را با دقت بالا شناسایی کنند.
- **پنهان‌کاری:**
  - با استفاده از سیگنال‌های ظریفی که اغلب نادیده گرفته می‌شوند، حملات کانال جانبی مبتنی بر هوش مصنوعی می‌توانند حتی در محیط‌های به‌خوبی ایمن‌شده نیز بدون تشخیص باقی بمانند.

۲-۳۳-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را آموزش می‌دهد تا الگوهای موجود در داده‌های کانال جانبی را تشخیص دهند و آن‌ها را به مقادیر مخفی مربوطه (مانند کلیدهای رمزنگاری) مرتبط کنند.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در داده‌های کانال جانبی بدون برچسب‌گذاری قبلی شناسایی می‌کند.
  - یادگیری تقویتی: استراتژی حمله را با پاداش دادن به بازیابی موفق کلید و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی، الگوهای پیچیده در داده‌های سری‌زمانی، مانند ردیابی‌های مصرف برق یا انتشارات الکترومغناطیسی، را تحلیل می‌کنند.
  - شبکه‌های عصبی بازگشتی وابستگی‌های ترتیبی در سیگنال‌های کانال جانبی را مدل‌سازی می‌کنند.
- **شبکه‌های مولد تخاصمی:**
  - داده‌های کانال جانبی مصنوعی اما واقع‌گرایانه برای اهداف آموزشی یا آزمایش استحکام اقدامات دفاعی ایجاد می‌کنند.
- **پردازش سیگنال:**
  - از تکنیک‌هایی مانند تبدیل‌های فوریه، تحلیل موجک یا تحلیل مؤلفه‌های اصلی (PCA) برای پیش‌پردازش و استخراج ویژگی‌های معنادار از داده‌های خام کانال جانبی استفاده می‌کند.
- **تحلیل سری‌های زمانی:**
  - از روش‌های آماری و یادگیری ماشین برای تحلیل وابستگی‌های زمانی در سیگنال‌های کانال

جانبی استفاده می‌کند و به پیش‌بینی رفتارهای آینده کمک می‌کند.

۳-۳۳-۲ دارایی‌های هدف

- **سیستم‌های رمزنگاری:**
  - ماژول‌های امنیتی سخت‌افزاری (HSMs)، کارت‌های هوشمند و دستگاه‌های تعبیه‌شده که الگوریتم‌های رمزنگاری را پیاده‌سازی می‌کنند.
- **خدمات ابری:**
  - ماشین‌های مجازی (VMs) و کانتینرهایی که منابع را در پلتفرم‌های چندمستاجره به اشتراک می‌گذارند، جایی که حملات مبتنی بر زمان‌بندی یا کش می‌توانند اطلاعات حساس را فاش کنند.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های اینترنت اشیا با منابع محدود که محافظت‌های کمی در برابر حملات کانال جانبی دارند.
- **سیستم‌های پرداخت:**
  - ترمینال‌های فروش، دستگاه‌های کارت‌خوان و سیستم‌های پرداخت بدون تماس که در برابر تحلیل مصرف برق یا نشت الکترومغناطیسی آسیب‌پذیر هستند.
- **سیستم‌های دولتی و نظامی:**
  - دستگاه‌های ارتباطی امن و تجهیزات رمزنگاری مورد استفاده در عملیات دفاعی و اطلاعاتی.
- **دستگاه‌های بهداشتی:**
  - ایمپلنت‌های پزشکی، مانیتورهای پوشیدنی سلامت و ابزارهای تشخیصی که داده‌های حساس بیماران را منتقل می‌کنند.

۴-۳۳-۲ آسیب‌پذیری‌ها

- **نشت سیگنال‌های فیزیکی:**
  - عملیات رمزنگاری سیگنال‌های قابل اندازه‌گیری مانند مصرف برق، تشعشعات الکترومغناطیسی یا تغییرات زمان‌بندی را منتشر می‌کنند.
- **اقدامات متقابل ضعیف:**
  - تکنیک‌های ناکافی مانند ماسک‌کردن، کورسازی یا تزریق نویز، سیستم‌ها را در برابر تحلیل کانال جانبی آسیب‌پذیر می‌کنند.

- **منابع مشترک:**
  - محیط‌های چندمستاجره، مانند پلتفرم‌های ابری، به مهاجمان امکان می‌دهند از کش‌ها، حافظه یا پردازنده‌های مشترک برای استنباط اطلاعات حساس سوءاستفاده کنند.
- **خطای انسانی:**
  - انتخاب‌های طراحی ضعیف، تنظیمات نادرست یا آزمایش‌های ناکافی در طول توسعه ممکن است آسیب‌پذیری‌هایی را ایجاد کند که از طریق حملات کانال جانبی قابل سوءاستفاده باشند.
- **عدم نظارت:**
  - عدم نظارت مداوم بر فعالیت‌های غیرعادی کانال جانبی، امکان پیشروی حملات بدون تشخیص را فراهم می‌کند.

۵-۳۳-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از سنسورها یا ابزارهای تخصصی برای ضبط سیگنال‌های کانال جانبی، مانند ردیابی‌های مصرف برق، انتشارات الکترومغناطیسی یا اندازه‌گیری‌های زمان‌بندی استفاده می‌کند.
- **پیش‌پردازش:**
  - مهاجم از تکنیک‌های پردازش سیگنال برای پاک‌سازی و نرمال‌سازی داده‌های جمع‌آوری‌شده استفاده کرده و نویز را حذف می‌کند و ویژگی‌های مرتبط را بهبود می‌بخشد.
- **آموزش مدل:**
  - مهاجم دل‌های یادگیری ماشین را بر روی مجموعه‌داده‌های برچسب‌دار آموزش داده تا سیگنال‌های کانال جانبی را با مقادیر مخفی (مانند کلیدهای رمزنگاری) مرتبط کنند.
- **اجرای حمله:**
  - مهاجم مدل آموزش‌دیده را برای تحلیل داده‌های کانال جانبی در زمان واقعی و استخراج اطلاعات حساس مستقر می‌کند.
- **پس‌پردازش:**
  - مهاجم داده‌های استخراج‌شده را با استفاده از تکنیک‌های پس‌پردازش بهبود داده تا دقت افزایش یابد و خطاها کاهش یابد.

- **سوءاستفاده:**
  - مهاجم از اسرار بازیابی‌شده برای اهداف مخرب، مانند رمزگشایی ارتباطات حساس یا جعل هویت کاربران قانونی استفاده می‌کند.

۲-۳۳-۶ نشانه‌های احتمال نفوذ

- **فعالیت غیرعادی سنسورها:**
  - وجود سنسورها یا ابزارهای نظارتی غیرمجاز در نزدیکی سیستم‌های حیاتی.
- **مصرف برق غیرعادی:**
  - انحراف در الگوهای مصرف برق که با عملیات یا رویدادهای خاص همبستگی دارد.
- **انتشارات الکترومغناطیسی غیرمنتظره:**
  - افزایش نویز الکترومغناطیسی یا تداخل در نزدیکی تجهیزات حساس.
- **ناهنجاری‌های زمان‌بندی:**
  - زمان‌های اجرای ناسازگار برای عملیات رمزنگاری یا سایر فرآیندهای بحرانی امنیتی.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول سیستم، مانند تلاش‌های مکرر برای دسترسی به منابع یا عملکردهای خاص.

۲-۳۳-۷ تأثیرات

- **نقض محرمانگی:**
  - افشای داده‌های حساس، مانند کلیدهای رمزنگاری، رمز عبور یا اطلاعات شخصی.
- **ضررهای مالی:**
  - دسترسی غیرمجاز به سیستم‌های مالی یا سرقت دارایی‌های دیجیتال.
- **اختلال در عملیات:**
  - اختلال در ارتباطات امن یا زیرساخت‌های حیاتی به دلیل به‌خطر افتادن سیستم‌های رمزنگاری.
- **آسیب به اعتبار:**
  - از دست دادن اعتماد مشتریان و ارزش برند پس از یک نقض.

- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، منجر به جریمه‌ها و پیامدهای قانونی می‌شود.
- **خطرات ایمنی:**
  - تهدید برای جان انسان‌ها در صورت به‌خطر افتادن دستگاه‌های پزشکی، سیستم‌های حمل‌ونقل یا کنترل‌های صنعتی.

۸-۳۳-۲ راه‌های مقابله

- **ماسک‌کردن و کورسازی:**
  - با افزودن تصادفی‌سازی به عملیات رمزنگاری، سیگنال‌های کانال جانبی را مبهم کنید و تحلیل آن‌ها را دشوارتر کنید.
- **تزریق نویز:**
  - نویز کنترل‌شده را به سیستم وارد کنید تا اندازه‌گیری‌های کانال جانبی مختل شود و کارایی آن‌ها کاهش یابد.
- **بهبود طراحی سخت‌افزار:**
  - از قطعات سخت‌افزاری تخصصی، مانند تراشه‌های مقاوم در برابر تحلیل مصرف برق (DPA)، برای به حداقل رساندن نشت استفاده کنید.
- **روش‌های کدنویسی امن:**
  - الگوریتم‌های ثابت‌زمانی را پیاده‌سازی کنید و از شاخه‌های شرطی که ممکن است اطلاعات زمان‌بندی را فاش کنند، اجتناب کنید.
- **بررسی‌های منظم:**
  - ممیزی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌های بالقوه کانال جانبی را شناسایی و برطرف کنید.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت کانال جانبی و ایجاد هشدار استفاده کنید.
- **جداسازی منابع:**
  - اطمینان حاصل کنید که منابع در محیط‌های چندمستاجره به‌درستی جدا شده‌اند تا از حملات cross-VM یا cross-container جلوگیری شود.

- **اقدامات امنیتی فیزیکی:**
  - سیستم‌های حیاتی را با موانع فیزیکی، محافظها یا محفظه‌های مقاوم در برابر دستکاری محافظت کنید تا دسترسی به سیگنال‌های کانال جانبی محدود شود.

۲-۳۴ بدافزارهای مبتنی بر هوش مصنوعی برای فرار از تشخیص

بدافزارهای مبتنی بر هوش مصنوعی که برای فرار از تشخیص طراحی شده‌اند[1]، به استفاده از هوش مصنوعی برای طراحی و استقرار بدافزارهایی اشاره می‌کنند که می‌توانند از سیستم‌های سنتی و پیشرفته تشخیص بدافزار، مانند نرم‌افزارهای آنتی‌ویروس، سیستم‌های تشخیص نفوذ (IDS)، محیط‌های sandbox و ابزارهای تحلیل رفتاری، عبور کنند. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند بدافزارهایی ایجاد کنند که رفتار، امضا یا الگوهای اجرایی خود را برای باقی ماندن در حالت غیرقابل تشخیص، حتی در محیط‌های بسیار امن، تطبیق می‌دهند.

این نوع حمله محدودیت‌های روش‌های تشخیص استاتیک و اکتشافی را مورد سوءاستفاده قرار می‌دهد و به مهاجمان اجازه می‌دهد تا بدون شناسایی شدن، به سیستم‌ها نفوذ کرده، داده‌ها را سرقت کنند یا عملیات را مختل کنند.

۲-۳۴-۱ ویژگی‌های کلیدی

- **رفتار پویا:**
  - هوش مصنوعی به بدافزار اجازه می‌دهد تا رفتار خود را بر اساس محیط تطبیق دهد و از تشخیص توسط ابزارهای امنیتی جلوگیری کند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین، فرآیند تولید انواع جدید بدافزار را خودکار می‌کنند، که باعث کاهش تلاش دستی و افزایش کارایی می‌شود.
- **انطباق‌پذیری:**
  - بدافزارهای مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و کد یا جریان اجرایی خود را برای فرار از دفاع‌های به‌روزرسانی‌شده، تغییر دهند.
- **پنهان‌کاری:**
  - با تقلید از رفتار نرم‌افزارهای قانونی یا رمزنگاری payloadها تا زمان اجرا، بدافزارهای مبتنی بر هوش مصنوعی در طول تحلیل پنهان می‌مانند.

1. Evading malware detection

- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان اجازه می‌دهد تا تعداد زیادی نمونه بدافزار منحصربه‌فرد تولید کنند و مکانیزم‌های تشخیص سنتی را تحت فشار قرار دهند.

۲-۳۴-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های برچسب‌خورده از بدافزارهای شناخته‌شده آموزش می‌دهد تا الگوها را یاد بگیرد و انواع جدیدی با امضاهای تغییر یافته ایجاد کند.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در کد بدافزار را شناسایی می‌کند تا تکنیک‌های بالقوه فرار را کشف کند.
  - یادگیری تقویتی: استراتژی‌های فرار را با پاداش دادن به موفقیت‌ها در اجتناب از تشخیص و جریمه کردن شکست‌ها، بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در کد باینری، ترافیک شبکه یا رفتار سیستم را تحلیل می‌کنند تا تکنیک‌های پنهان‌سازی را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: نمونه‌های بدافزار مصنوعی اما واقع‌گرایانه برای اهداف آموزشی یا آزمایش اقدامات دفاعی ایجاد می‌کنند.
- **جهش کد:**
  - از هوش مصنوعی برای تغییر تصادفی کد بدافزار در حالی که عملکرد آن حفظ می‌شود، استفاده می‌کند و انواع چندشکلی یا متامورفیک ایجاد می‌کند.
- **تقلید رفتاری:**
  - از هوش مصنوعی برای شبیه‌سازی رفتار نرم‌افزارهای قانونی استفاده می‌کند، که تشخیص بدافزار توسط ابزارهای تحلیل رفتاری را سخت‌تر می‌کند.
- **فرار از sandbox:**
  - از هوش مصنوعی برای تشخیص و اجتناب از محیط‌های sandbox استفاده می‌کند، تا اطمینان حاصل شود که بدافزار فقط در شرایط واقعی اجرا می‌شود.

۳-۳۴-۲ دارایی‌های هدف

- **دستگاه‌های نهایی:**
  - کامپیوترها، گوشی‌های هوشمند، تبلت‌ها و دستگاه‌های IoT که می‌توانند هدف بدافزار قرار گیرند تا داده‌ها را سرقت کنند یا عملیات را مختل کنند.
- **سیستم‌های شبکه:**
  - شبکه‌های سازمانی، پلتفرم‌های ابری یا سیستم‌های کنترل صنعتی که در برابر انتشار پنهانی بدافزار آسیب‌پذیر هستند.
- **راه‌حل‌های آنتی‌ویروس:**
  - نرم‌افزارهای آنتی‌ویروس سنتی که به تشخیص مبتنی بر امضا یا تحلیل استاتیک متکی هستند و می‌توانند توسط بدافزارهای تولیدشده توسط هوش مصنوعی دور زده شوند.
- **محیط‌های sandbox:**
  - sandboxهای امنیتی که برای تحلیل فایل‌های مشکوک استفاده می‌شوند و می‌توانند توسط بدافزارهای مبتنی بر هوش مصنوعی که در طول تست بی‌ضرر رفتار می‌کنند، فریب بخورند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تاسیسات بهداشتی، موسسات مالی یا سیستم‌های حمل‌ونقل که بدافزارهای غیرقابل تشخیص می‌توانند آسیب‌های قابل توجهی ایجاد کنند.

۴-۳۴-۲ آسیب‌پذیری‌ها

- **تشخیص مبتنی بر امضا:**
  - راه‌حل‌های آنتی‌ویروس که از امضاهای قدیمی یا ناقص استفاده می‌کنند، به راحتی توسط بدافزارهای چندشکلی یا متامورفیک دور زده می‌شوند.
- **محدودیت‌های تحلیل استاتیک:**
  - عدم توانایی در تحلیل پویا، تشخیص بدافزارهای مبتنی بر هوش مصنوعی را که فقط در زمان اجرا ماهیت واقعی خود را نشان می‌دهند، دشوار می‌کند.
- **خلأهای نظارت رفتاری:**
  - ابزارهای ناکافی تحلیل رفتاری، انحرافات ظریف ناشی از بدافزارهای پنهان‌شده توسط هوش مصنوعی را تشخیص نمی‌دهند.

- **خطای انسانی:**
  - اشتباهات توسعه‌دهندگان، مدیران یا اپراتورها به طور ناخواسته آسیب‌پذیری‌هایی را آشکار می‌کنند که توسط بدافزارهای تطبیقی مورد سوءاستفاده قرار می‌گیرند.
- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - نقص‌های کشف‌نشده در نرم‌افزار یا سخت‌افزار، نقاط ورودی برای بدافزارهای مبتنی بر هوش مصنوعی فراهم می‌کنند تا جای پای خود را محکم کنند.

۵-۳۴-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره سیستم‌های هدف، از جمله دفاع‌ها، پیکربندی‌ها و رفتارهای معمول استفاده می‌کند.
- **طراحی بدافزار:**
  - مهاجم از مدل‌های یادگیری ماشین برای تحلیل نمونه‌های موجود بدافزار و طراحی انواع جدیدی که قادر به فرار از تشخیص هستند، استفاده می‌کند.
- **تکنیک‌های پنهان‌سازی:**
  - از روش‌های پنهان‌سازی مبتنی بر هوش مصنوعی، مانند رمزنگاری، بسته‌بندی یا تقلید رفتاری، برای پنهان کردن ماهیت واقعی بدافزار استفاده می‌شود.
- **تست و بهبود:**
  - بدافزار پنهان‌شده در محیط‌های شبیه‌سازی‌شده تست می‌شود تا اطمینان حاصل شود که از اقدامات امنیتی عبور می‌کند و رفتار آن بر این اساس بهبود می‌یابد.
- **استقرار:**
  - بدافزار از طریق ایمیل‌های فیشینگ، کیت‌های اکسپلویت یا سایر روش‌ها به سیستم هدف تحویل داده می‌شود.
- **ماندگاری:**
  - دسترسی به سیستم‌های آلوده شده را با تکامل مداوم بدافزار برای تطبیق با دفاع‌های در حال تغییر، حفظ می‌شود.

۲-۳۴-۶ نشانه‌های احتمال نفوذ

- **تغییرات غیرمعمول فایل:**
  - ناهنجاری‌ها در الگوهای ایجاد، حذف یا تغییر فایل‌ها که با فعالیت‌های قانونی مرتبط نیستند.
- **افزایش استفاده از منابع:**
  - فعالیت بیشتر CPU، حافظه یا I/O دیسک به دلیل اجرای بدافزار پنهان‌شده در پس‌زمینه.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول سیستم، مانند اتصالات شبکه غیرمنتظره یا افزایش امتیاز دسترسی غیرمجاز.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به احراز هویت ناموفق، فرآیندهای غیرمعمول یا فعالیت‌های غیرمنتظره.
- **ترافیک شبکه غیرمنتظره:**
  - انحرافات در ترافیک خروجی، مانند اتصالات به آدرس‌های IP یا دامنه‌های ناشناخته.

۲-۳۴-۷ تأثیرات

- **نشت داده:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به زیان‌های مالی یا آسیب به اعتبار می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل سرقت یا نقض شدن اعتبارنامه‌ها یا سوءاستفاده از آسیب‌پذیری‌ها.
- **زیان‌های مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزار یا اختلال در خدمات که باعث آسیب مالی می‌شود.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا ریزبرنامه‌های حیاتی، که سیستم‌ها را غیرقابل استفاده می‌کند.

۸-۳۴-۲ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - از راه‌حل‌های آنتی‌ویروس مبتنی بر هوش مصنوعی که روی مجموعه‌داده‌های بزرگی از نمونه‌های بدافزار آموزش دیده‌اند، استفاده کنید تا تهدیدات پنهان‌شده را به‌طور موثرتری شناسایی کنند.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص فعالیت‌های غیرعادی که نشان‌دهنده بدافزارهای فرار هستند، استفاده کنید.
- **تحلیل پویا:**
  - محیط‌های sandbox را پیاده‌سازی کنید که قادر به اجرا و تحلیل فایل‌های مشکوک در تنظیمات کنترل‌شده باشند تا رفتارهای پنهان را آشکار کنند.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، ریزبرنامه‌ها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده بدافزارهای فرار را برطرف کنید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش بدافزار محدود شود و تأثیر عفونت‌های موفق کاهش یابد.
- **بهداشت رمزنگاری:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حال انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده به حداقل برسد.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد علائم عفونت بدافزار، مانند تلاش‌های فیشینگ یا رفتار غیرمعمول سیستم، آموزش دهید.

- **هوش‌مندی تهدید پیشرفته:**
  - از پلتفرم‌های هوش‌مندی تهدید مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور فرار و بدافزارهای تطبیقی مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حملات بدافزارهای فرار را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.

۲-۳۵ دور زدن سیستم‌های تشخیص تهدیدات داخلی

سیستم‌های تشخیص تهدیدات داخلی برای شناسایی و کاهش خطرات ناشی از افرادی طراحی شده‌اند که ممکن است از دسترسی‌های خود در یک سازمان برای اهداف مخرب سوءاستفاده کنند. در یک حمله مبتنی بر هوش مصنوعی برای دور زدن سیستم‌های تشخیص تهدیدات داخلی، از هوش مصنوعی برای تحلیل و دستکاری رفتار کاربران داخلی یا حساب‌های به‌خطرافتاده به‌گونه‌ای استفاده می‌شود که مکانیزم‌های سنتی و پیشرفته تشخیص تهدیدات داخلی را دور می‌زنند. با استفاده از یادگیری ماشین و تحلیل رفتاری، مهاجمان می‌توانند فعالیت‌های قانونی کاربران را شبیه‌سازی کنند، اقدامات مشکوک را پنهان کنند یا نقاط کور در الگوریتم‌های تشخیص را مورد سوءاستفاده قرار دهند.

دور زدن سیستم‌های تشخیص تهدیدات داخلی با استفاده از هوش مصنوعی نشان‌دهنده تحولی قابل توجه در تهدیدات سایبری است که از قدرت یادگیری ماشین و تحلیل رفتاری برای دور زدن حتی پیشرفته‌ترین سیستم‌های تشخیص استفاده می‌کند. این نوع حمله خطرناک است زیرا اعتماد به سیستم‌های داخلی و پرسنل را تضعیف می‌کند و به مهاجمان امکان می‌دهد تا داده‌های حساس را خارج کنند، عملیات را مختل کنند یا آسیب‌های اعتباری ایجاد کنند بدون اینکه شناسایی شوند.

۲-۳۵-۱ ویژگی‌های کلیدی

- **تقلید رفتاری:**
  - هوش مصنوعی رفتار عادی کاربران را تحلیل می‌کند تا الگوها را تکرار کند و فعالیت‌های مخرب را قانونی جلوه دهد.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند شناسایی آسیب‌پذیری‌ها در سیستم‌های تشخیص و ایجاد استراتژی‌های دور زدن را به‌صورت خودکار انجام می‌دهند.
- **انطباق‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در

قوانین تشخیص، آستانه‌ها یا شرایط محیطی سازگار شوند.

- **پنهان‌کاری:**
  - با ترکیب شدن در فعالیت‌های عادی یا سوءاستفاده از فاصله‌های زمانی، هوش مصنوعی اطمینان می‌دهد که اقدامات مخرب شناسایی نشوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا رفتارها یا توالی‌های خاصی را که از تشخیص فرار می‌کنند، شناسایی کنند و در عین حال به اهداف خود دست یابند.

۲-۳۵-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های فعالیت‌های قانونی و مخرب کاربران آموزش می‌دهد تا الگوها را یاد بگیرد و تکنیک‌های دور زدن را بهبود بخشد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در رفتار کاربران را شناسایی می‌کند تا نقاط ضعف احتمالی در سیستم‌های تشخیص را آشکار کند.
  - یادگیری تقویتی: استراتژی‌های دور زدن را با پاداش دادن به موفقیت‌های اجتناب از تشخیص و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، وابستگی‌های زمانی پیچیده در رفتار کاربران را تحلیل می‌کنند تا اقدامات آینده را پیش‌بینی و تقلید کنند.
  - شبکه‌های مولد تخاصمی: لاگ‌های فعالیت کاربران مصنوعی اما واقع‌گرایانه ایجاد می‌کنند که برای اهداف آموزشی یا تست اقدامات دفاعی استفاده می‌شوند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربران استفاده می‌کند تا اقداماتی را طراحی کند که با الگوهای مورد انتظار هماهنگ باشند.
- **دور زدن تشخیص ناهنجاری:**
  - از هوش مصنوعی برای شناسایی و سوءاستفاده از آستانه‌ها یا مرزها در سیستم‌های تشخیص ناهنجاری استفاده می‌کند تا اطمینان حاصل کند که اقدامات مخرب زیر سطح سوءظن باقی می‌مانند.

- **دستکاری زمانی:**
  - از هوش مصنوعی برای توزیع یا تصادفی‌سازی اقدامات مخرب در طول زمان استفاده می‌کند تا از تشخیص توسط سیستم‌هایی که بر ناهنجاری‌های کوتاه‌مدت متمرکز هستند، جلوگیری کند.

۳-۳۵-۲ دارایی‌های هدف

- **سیستم‌های سازمانی:**
  - شبکه‌های شرکتی، سرورهای فایل و سیستم‌های پایگاه‌داده که داده‌های حیاتی کسب‌وکار را ذخیره می‌کنند.
- **حساب‌های کاربری:**
  - حساب‌های کارمندان یا پیمانکاران با دسترسی‌های بالا یا دسترسی به اطلاعات حساس.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تأسیسات بهداشتی، مؤسسات مالی یا سیستم‌های حمل‌ونقل که تهدیدات داخلی می‌توانند آسیب‌های قابل توجهی ایجاد کنند.
- **خدمات ابری:**
  - برنامه‌ها، ذخیره‌سازی یا منابع محاسباتی میزبانی‌شده در ابر که از طریق اعتبارنامه‌های داخلی به‌خطرافتاده قابل دسترسی هستند.
- **اجزای زنجیره تأمین:**
  - فروشندگان، شرکا یا تأمین‌کنندگان شخص ثالث که حساب‌های آن‌ها ممکن است به‌عنوان نقاط ورودی برای حملات داخلی مورد استفاده قرار گیرند.

۴-۳۵-۲ آسیب‌پذیری‌ها

- **تحلیل رفتاری ضعیف:**
  - ابزارهای تحلیل رفتاری که به‌اندازه کافی آموزش ندیده‌اند یا قدیمی هستند، نمی‌توانند انحرافات ظریف ناشی از تکنیک‌های دور زدن مبتنی بر هوش مصنوعی را تشخیص دهند.
- **خطای انسانی:**
  - سیستم‌های تشخیص نادرست پیکربندی‌شده یا نظارت ناکافی، سازمان‌ها را در برابر تهدیدات داخلی پیچیده آسیب‌پذیر می‌کنند.

- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های مبتنی بر قوانین ثابت یا تشخیص مبتنی بر امضا، موفقیت حملات مبتنی بر هوش مصنوعی را آسان‌تر می‌کند.
- **اتکای بیش از حد به خودکارسازی:**
  - اتکای بیش از حد به سیستم‌های تشخیص خودکار بدون نظارت انسانی می‌تواند منجر به از دست رفتن فرصت‌های شناسایی تهدیدات پیچیده شود.
- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - نقص‌های کشف‌نشده در الگوریتم‌های تشخیص یا پیکربندی سیستم‌ها، نقاط ورودی برای حملات داخلی مبتنی بر هوش مصنوعی فراهم می‌کنند.

۵-۳۵-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره کاربران هدف، از جمله نقش‌ها، مجوزها و رفتار معمول آن‌ها استفاده می‌شود.
- **تحلیل و آموزش:**
  - مدل‌های یادگیری ماشین برای تحلیل فعالیت‌های تاریخی کاربران و آموزش روی الگوهای رفتار قانونی مستقر می‌شوند.
- **توسعه استراتژی دور زدن:**
  - اقداماتی طراحی می‌شود که با رفتار عادی کاربران هماهنگ باشد و در عین حال اهداف مهاجمان، مانند خروج داده‌ها یا افزایش دسترسی، را پیش ببرد.
- **اجرا:**
  - اقدامات مخرب به‌گونه‌ای انجام می‌شود که فعالیت‌های قانونی کاربران را تقلید کند و اطمینان حاصل شود که توسط سیستم‌های تشخیص تهدیدات داخلی شناسایی نمی‌شوند.
- **انطباق:**
  - بازخورد سیستم‌های تشخیص به‌طور مداوم نظارت می‌شود و استراتژی‌ها در زمان واقعی تنظیم می‌شوند تا از ایجاد هشدار جلوگیری شود.
- **پایداری:**
  - دسترسی به حساب‌های دارای دسترسی بالا یا سیستم‌ها با بهبود مستمر تکنیک‌های دور

زدن بر اساس نتایج مشاهده‌شده حفظ می‌شود.

۲-۳۵-۶ شانه‌های احتمال نفوذ

- **ناهنجاری‌های ظریف:**
  - انحرافات جزئی در رفتار کاربران، مانند زمان‌های ورود غیرمعمول یا افزایش جزئی در حجم دسترسی به داده‌ها.
- **الگوهای دسترسی غیرمنتظره:**
  - اقداماتی که با رفتار معمول کاربران ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره یا تلاش‌های دسترسی غیرمجاز.
- **استفاده از منابع:**
  - استفاده بیشتر از CPU، حافظه یا شبکه که با عملیات قانونی مرتبط نیست.
- **خروج داده‌ها:**
  - الگوهای ترافیک خروجی غیرمعمول، مانند اتصالات به آدرس‌های IP یا دامنه‌های ناشناخته.

۲-۳۵-۷ تأثیرات

- **نشت داده‌ها:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به ضررهای مالی یا آسیب‌های شهرت می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به‌دلیل حساب‌های داخلی به‌خطرافتاده.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلال در خدمات که باعث آسیب‌های مالی می‌شوند.
- **جریمه‌های نظارتی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، که منجر به جریمه‌ها و اقدامات قانونی می‌شود.

- **فرسایش اعتماد:**
  - تضعیف اعتماد به سیستم‌های داخلی، کارمندان یا فرآیندها، که منجر به چالش‌های بلندمدت سازمانی می‌شود

۸-۳۵-۲ راه‌های مقابله

- **تحلیل رفتاری پیشرفته:**
  - از تحلیل رفتاری مبتنی بر هوش مصنوعی استفاده کنید تا انحرافات ظریف در فعالیت کاربران که نشان‌دهنده تهدیدات داخلی هستند، شناسایی شوند.
- **نظارت مداوم:**
  - راه‌حل‌های نظارت در زمان واقعی پیاده‌سازی کنید که قادر به تحلیل رفتارهای فردی و جمعی کاربران برای شناسایی ناهنجاری‌ها باشند.
- **مدیریت دسترسی‌های ممتاز:**
  - کنترل‌های دسترسی سخت‌گیرانه، احراز هویت چندعاملی و اصل حداقل دسترسی را اعمال کنید تا تأثیر حساب‌های به‌خطرافتاده محدود شود.
- **تشخیص مبتنی بر هوش مصنوعی:**
  - از سیستم‌های تشخیص مبتنی بر یادگیری ماشین استفاده کنید که روی مجموعه‌داده‌های بزرگی از فعالیت‌های قانونی و مخرب آموزش دیده‌اند تا دقت را بهبود بخشند.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **ثبت فعالیت کاربران:**
  - گزارش‌های دقیقی از تمام فعالیت‌های کاربران نگه دارید و مکانیزم‌های تشخیص ناهنجاری را پیاده‌سازی کنید تا اقدامات مشکوک را علامت‌گذاری کنند.
- **آموزش و آگاهی:**
  - کارمندان را در مورد شناسایی نشانه‌های تهدیدات داخلی و رعایت بهداشت سایبری آموزش دهید.
- **هوشمندسازی تهدیدات پیشگیرانه:**
  - از پلتفرم‌های هوشمند مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور دور

زدن و تهدیدات تطبیقی مطلع شوید.

- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی‌ای توسعه دهید و پیاده‌سازی کنید تا حملات داخلی را به‌سرعت شناسایی، مهار و کاهش دهید.
- **کنترل غیرمتمرکز:**
  - مسئولیت‌های کنترل و نظارت را در چندین لایه توزیع کنید تا وابستگی به هر نقطه شکست واحد کاهش یابد.

۲-۳۶ دور زدن فیلترهای ایمیل

سیستم‌های فیلتر ایمیل برای شناسایی و مسدود کردن ایمیل‌های مخرب مانند فیشینگ، اسپم یا پیام‌های حاوی بدافزار طراحی شده‌اند. در یک حمله مبتنی بر هوش مصنوعی برای دور زدن فیلترهای ایمیل، از هوش مصنوعی برای ایجاد و ارسال ایمیل‌هایی استفاده می‌شود که مکانیزم‌های سنتی و پیشرفته فیلتر کردن را دور می‌زنند. با استفاده از یادگیری ماشین و پردازش زبان طبیعی، مهاجمان می‌توانند محتوای ایمیل‌های بسیار متقاعدکننده، آگاه از زمینه و تطبیق‌پذیر ایجاد کنند که از تشخیص فرار کرده و در عین حال به هدف خود دست یابند.

این نوع حمله می‌تواند اثربخشی راه‌حل‌های امنیتی ایمیل را تضعیف می‌کند و به مهاجمان امکان می‌دهد تا محموله‌های مخرب را تحویل دهند، اعتبار را سرقت کنند یا اطلاعات نادرست را منتشر کنند.

۲-۳۶-۱ ویژگی‌های کلیدی

- **محتوای بسیار شخصی‌سازی‌شده:**
  - هوش مصنوعی ایمیل‌هایی ایجاد می‌کند که برای گیرندگان خاص تنظیم شده‌اند، که احتمال تعامل را افزایش داده و سوءظن را کاهش می‌دهد.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند ایجاد، ارسال و بهبود ایمیل‌ها را به‌صورت خودکار انجام می‌دهند، که تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **انطباق‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در قوانین فیلتر، الگوریتم‌ها یا شرایط محیطی سازگار شوند.
- **پنهان‌کاری:**
  - با تقلید از سبک‌های ارتباطی قانونی و اجتناب از محرک‌های شناخته‌شده اسپم، هوش

مصنوعی اطمینان می‌دهد که ایمیل‌ها شناسایی نشوند.

- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا حجم زیادی از ایمیل‌های سفارشی‌شده را به‌طور همزمان به چندین هدف ارسال کنند.

۲-۳۶-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **پردازش زبان طبیعی:**
  - تولید متن: از مدل‌های مبتنی بر ترنسفورمر مانند GPT-3 یا BERT برای ایجاد محتوای ایمیل منسجم و مرتبط با زمینه استفاده می‌کند.
  - تحلیل احساسات: ترجیحات و احساسات گیرنده را تحلیل می‌کند تا خطوط موضوعی و متن بدنه متقاعدکننده ایجاد کند.
  - مدل‌سازی موضوعی: موضوعات یا تم‌های داغ را شناسایی می‌کند تا محتوای ایمیل را با علایق فعلی هماهنگ کند.
- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های کمپین‌های ایمیل موفق و ناموفق آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا ناهنجاری‌ها در ساختار ایمیل‌ها را شناسایی می‌کند تا نقاط ضعف احتمالی در سیستم‌های فیلتر کردن را آشکار کند.
  - یادگیری تقویتی: تولید ایمیل را با پاداش دادن به تحویل موفق و جریمه کردن تلاش‌های مسدود شده بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و شبکه‌های حافظه طولانی کوتاه-مدت (LSTM)، وابستگی‌های زمانی در مکالمات ایمیل را تحلیل می‌کنند تا واقع‌گرایی را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: قالب‌های ایمیل مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا تکنیک‌های دور زدن را تست و بهبود بخشند.
- **تقلید رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربران استفاده می‌کند تا الگوهای ارتباطی قانونی را تقلید کند.

- **جعل دامنه:**
  - از هوش مصنوعی برای ایجاد نام‌های دامنه یا هویت‌های فرستنده که قانونی به نظر می‌رسند استفاده می‌کند تا از تشخیص توسط سیستم‌های فیلتر مبتنی بر دامنه جلوگیری کند.

۲-۳۶-۳ دارایی‌های هدف

- **ایمیل‌های شرکتی:**
  - حساب‌های ایمیل کسب‌وکار که داده‌های حساس، اعتبار یا اطلاعات مالی را ذخیره می‌کنند.
- **حساب‌های شخصی:**
  - حساب‌های ایمیل فردی که در برابر فیشینگ، سرقت اعتبار یا مهندسی اجتماعی آسیب‌پذیر هستند.
- **مؤسسات مالی:**
  - بانک‌ها، اتحادیه‌های اعتباری یا پردازنده‌های پرداخت که کلاهبرداری مبتنی بر ایمیل می‌تواند منجر به ضررهای قابل توجهی شود.
- **آژانس‌های دولتی:**
  - اهداف با ارزش بالا که اطلاعات طبقه‌بندی‌شده را ذخیره می‌کنند یا زیرساخت‌های حیاتی را کنترل می‌کنند.
- **خدمات ابری:**
  - پلتفرم‌های ایمیل میزبانی‌شده در ابر که حساب‌های به‌خطرافتاده می‌توانند برای حملات بیشتر مورد استفاده قرار گیرند.

۲-۳۶-۴ آسیب‌پذیری‌ها

- **فیلترهای اسپم ضعیف:**
  - فیلترهای سنتی مبتنی بر قوانین یا امضا نمی‌توانند با پیچیدگی ایمیل‌های تولیدشده توسط هوش مصنوعی همگام شوند.
- **نظارت رفتاری ناکافی:**
  - فقدان ابزارهای قوی تحلیل رفتاری، انحرافات ظریف ناشی از ایمیل‌های مبتنی بر هوش مصنوعی را تشخیص نمی‌دهد.

- **خطای انسانی:**
  - کارمندان یا کاربران ممکن است به دلیل قانونی به نظر رسیدن ایمیل‌های بسیار شخصی‌سازی‌شده و متقاعدکننده، فریب بخورند.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های تشخیص قدیمی یا نادرست پیاده‌سازی‌شده، سازمان‌ها را در برابر تهدیدات ایمیل تطبیقی آسیب‌پذیر می‌کند.
- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - نقص‌های کشف‌نشده در الگوریتم‌های فیلتر ایمیل یا پیکربندی‌ها، نقاط ورودی برای حملات مبتنی بر هوش مصنوعی فراهم می‌کنند.

۵-۳۶-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره گیرندگان هدف، از جمله نقش‌ها، ترجیحات و فعالیت‌های اخیر آن‌ها استفاده می‌شود.
- **تولید محتوا:**
  - مدل‌های پردازش زبان طبیعی برای ایجاد محتوای ایمیل شخصی‌سازی‌شده و آگاه از زمینه که برای هر گیرنده تنظیم شده‌اند، مستقر می‌شوند.
- **ایجاد هویت فرستنده:**
  - از هوش مصنوعی برای ایجاد هویت‌های فرستنده متقاعدکننده، نام‌های دامنه یا امضاهای ایمیل استفاده می‌شود که از تشخیص جعل دامنه جلوگیری می‌کند.
- **تست و بهبود:**
  - ایمیل‌های تولیدشده را در برابر سیستم‌های فیلتر شبیه‌سازی‌شده تست می‌کنند تا اطمینان حاصل شود که از تشخیص فرار می‌کنند و محتوا را بر این اساس بهبود می‌بخشند.
- **استقرار:**
  - ایمیل‌های طراحی‌شده را به صندوق‌های ورودی هدف تحویل می‌دهند، اطمینان حاصل می‌کنند که قانونی به نظر می‌رسند و تعامل را تشویق می‌کنند.
- **پایداری:**
  - محتوای ایمیل و تاکتیک‌ها را به‌طور مداوم بر اساس بازخورد از تلاش‌های موفق به‌روزرسانی

می‌کنند تا قابلیت‌های دور زدن بلندمدت حفظ شود.

۲-۳۶-۶ نشانه‌های احتمال نفوذ

- **رفتار غیرعادی فرستنده:**
  - ناهنجاری‌ها در دامنه‌های فرستنده، هدرهای ایمیل یا فراداده‌ها که از الگوهای معمول منحرف می‌شوند.
- **لینک‌ها یا پیوست‌های مشکوک:**
  - ایمیل‌های حاوی URLهای کوتاه‌شده، انواع فایل ناشناخته یا درخواست‌های اطلاعات حساس.
- **درخواست‌های غیرمنتظره:**
  - درخواست‌های فوری یا خارج از شخصیت برای پول، اعتبار یا داده‌های محرمانه.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سرور ایمیل مربوط به تحویل‌های ناموفق، فعالیت غیرمنتظره یا تلاش‌های دسترسی غیرمجاز.
- **بازخورد گیرنده:**
  - گزارش‌های کاربران که ایمیل‌های مشکوکی را که ممکن است از سیستم‌های فیلتر عبور کرده‌اند، علامت‌گذاری می‌کنند.

۲-۳۶-۷ تأثیرات

- **حملات فیشینگ:**
  - تحویل موفقیت‌آمیز ایمیل‌های فیشینگ که منجر به سرقت اعتبار، عفونت‌های باج‌افزاری یا کلاهبرداری مالی می‌شود.
- **نشت داده‌ها:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به ضررهای مالی یا آسیب‌های اعتباری می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به‌دلیل حساب‌های به‌خطرافتاده یا آسیب‌پذیری‌های مورد سوءاستفاده قرار گرفته.

- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، انتقال‌های بانکی جعلی یا اختلال در خدمات که باعث آسیب‌های مالی می‌شوند.
- **آسیب به شهرت:**
  - از دست دادن اعتماد مشتریان و ارزش برند پس از یک نفوذ یا سوءاستفاده از اعتبارهای سرقت‌شده.

۸-۳۶-۲ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های فیلتر ایمیل مبتنی بر هوش مصنوعی را مستقر کنید که روی مجموعه‌داده‌های بزرگی از ایمیل‌های قانونی و مخرب آموزش دیده‌اند تا دقت تشخیص را بهبود بخشند.
- **تحلیل رفتاری:**
  - از تحلیل رفتاری مبتنی بر هوش مصنوعی استفاده کنید تا فعالیت‌های غیرعادی ایمیل که نشان‌دهنده دور زدن مبتنی بر هوش مصنوعی هستند، حتی اگر محتوا قانونی به نظر برسد، شناسایی شوند.
- **هوشمندسازی تهدیدات پیشرفته:**
  - از پلتفرم‌های هوشمند مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور دور زدن و تهدیدات تطبیقی مطلع شوید.
- **دفاع چندلایه:**
  - مکانیزم‌های فیلتر سنتی را با تشخیص ناهنجاری مبتنی بر هوش مصنوعی، sandboxing و نظارت در زمان واقعی ترکیب کنید تا محافظت جامع فراهم شود.
- **به‌روزرسانی‌های منظم:**
  - تمام سیستم‌های فیلتر ایمیل، فرمور و پیکربندی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده قرار گرفته توسط حملات مبتنی بر هوش مصنوعی را برطرف کنید.
- **آموزش کاربران:**
  - کارمندان و کاربران را در مورد شناسایی نشانه‌های فیشینگ، لینک‌های مشکوک یا درخواست‌های غیرمعمول آموزش دهید تا فرهنگ هوشیاری را تقویت کنند.

- **احراز هویت دامنه:**
  - SPF (چارچوب سیاست فرستنده)، DKIM (پست الکترونیکی شناسایی‌شده با کلید دامنه) و DMARC (احراز هویت پیام مبتنی بر دامنه، گزارش و انطباق) را پیاده‌سازی کنید تا اصالت فرستنده را تأیید کنید.
- **لیست سیاه پیش‌گیرانه:**
  - از تحلیل پیش‌بینانه استفاده کنید تا دامنه‌ها یا هویت‌های فرستنده احتمالاً مخرب را بر اساس الگوهای مشاهده‌شده پیش‌گیرانه لیست سیاه کنید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی‌ای توسعه دهید و پیاده‌سازی کنید تا حملات ایمیل موفق را به‌سرعت شناسایی، مهار و کاهش دهید.
- **رمزنگاری قوی:**
  - ارتباطات و داده‌های حساس را رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده را به حداقل برسانید.

۲-۳۷ فرار از سیستم تشخیص نفوذ شبکه (IDS)

سیستم تشخیص نفوذ شبکه (IDS) یک ابزار امنیتی است که برای نظارت و تحلیل ترافیک شبکه به منظور شناسایی فعالیت‌های مخرب، مانند دسترسی غیرمجاز، استخراج داده یا حملات طراحی شده است. در حمله فرار از IDS شبکه مبتنی بر هوش مصنوعی، از هوش مصنوعی برای ایجاد و استقرار الگوهای ترافیکی استفاده می‌شود که از سیستم‌های IDS سنتی و پیشرفته عبور می‌کنند. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند نحوه عملکرد IDS را تحلیل کنند، نقاط ضعف آن را شناسایی کنند و ترافیکی ایجاد کنند که رفتار قانونی را تقلید می‌کند در حالی که اقدامات مخرب را انجام می‌دهد.

این نوع حمله اثربخشی IDS را تضعیف می‌کند و به مهاجمان امکان می‌دهد بدون تشخیص، به شبکه‌ها نفوذ کنند، داده‌ها را سرقت کنند یا عملیات را مختل کنند.

۲-۳۷-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند شناسایی آسیب‌پذیری‌ها در IDS، ایجاد تکنیک‌های فرار و اجرای حملات را خودکار می‌کند و تلاش دستی را کاهش می‌دهد.

- **انعطاف‌پذیری:**
  - فرارکننده‌های مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در پیکربندی IDS، الگوریتم‌ها یا شرایط محیطی سازگار شوند.
- **پنهان‌کاری:**
  - با تقلید از الگوهای ترافیکی قانونی یا رمزنگاری payloadها تا زمان اجرا، هوش مصنوعی اطمینان می‌دهد که فعالیت‌های مخرب بدون تشخیص باقی بمانند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند قوانین یا آستانه‌های خاص IDS را برای عبور بدون ایجاد هشدار شناسایی کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد حجم زیادی از ترافیک پنهان را در چندین هدف به‌طور همزمان تولید و توزیع کنند.

۲-۳۷-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های ترافیک قانونی و مخرب آموزش می‌دهد تا الگوها را یاد بگیرد و استراتژی‌های فرار را بهبود بخشد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در الگوهای ترافیک شناسایی می‌کند تا نقاط کور احتمالی در پیکربندی IDS را آشکار کند.
  - یادگیری تقویتی: استراتژی‌های فرار را با پاداش دادن به موفقیت‌های اجتناب از تشخیص و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، وابستگی‌های زمانی در ترافیک را تحلیل می‌کنند تا واقع‌گرایی را بهبود بخشند و از تشخیص مبتنی بر ناهنجاری فرار کنند.
  - شبکه‌های مولد تخاصمی: الگوهای ترافیکی مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های فرار را آزمایش و بهبود بخشند.
- **تقلید رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار طبیعی کاربر یا سیستم استفاده

می‌کند تا مهاجمان بتوانند فعالیت‌های مخرب را با فعالیت‌های قانونی ترکیب کنند.

- **مبهم‌سازی ترافیک:**
  - از هوش مصنوعی برای قطعه‌قطعه کردن یا رمزنگاری ترافیک مخرب استفاده می‌کند تا در طول تحلیل بی‌خطر به نظر برسد.
- **اجتناب از امضا:**
  - پایگاه‌داده‌های امضای IDS را تحلیل می‌کند تا ترافیکی ایجاد کند که از تطابق با الگوهای حمله شناخته‌شده اجتناب کند.

۲-۳۷-۳ دارایی‌های هدف

- **شبکه‌های سازمانی:**
  - شبکه‌های شرکتی که داده‌های حساس، مالکیت فکری یا اطلاعات مالی در آن‌ها ذخیره می‌شود.
- **خدمات ابری:**
  - پلتفرم‌های ابری که برنامه‌ها، ذخیره‌سازی یا منابع محاسباتی را میزبانی می‌کنند و از طریق اعتبارنامه‌های نفوذشده قابل دسترسی هستند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، مراکز بهداشتی، مؤسسات مالی یا شبکه‌های حمل‌ونقل که نفوذهای بدون تشخیص می‌توانند آسیب‌های جدی ایجاد کنند.
- **شبکه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های اینترنت اشیا با منابع محدود که اغلب به عنوان نقاط ورودی برای کمپین‌های بزرگتر استفاده می‌شوند.
- **سازمان‌های دولتی:**
  - اهداف باارزش که اطلاعات طبقه‌بندی‌شده را ذخیره می‌کنند یا سیستم‌های امنیت ملی را کنترل می‌کنند.

۲-۳۷-۴ آسیب‌پذیری‌ها

- **تشخیص ناهنجاری ضعیف:**
  - IDSهایی که به الگوریتم‌های تشخیص ناهنجاری قدیمی یا ناقص متکی هستند، قادر به تشخیص انحرافات ظریف ناشی از ترافیک فرار مبتنی بر هوش مصنوعی نیستند.

- **خطای انسانی:**
  - قوانین نادرست پیکربندی‌شده IDS، اعتبارنامه‌های مشترک یا آموزش ناکافی به‌طور ناخواسته آسیب‌پذیری‌هایی را در معرض سوءاستفاده توسط تکنیک‌های فرار قرار می‌دهند.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های قدیمی یا دفاع‌های نادرست پیاده‌سازی‌شده، سازمان‌ها را در برابر استراتژی‌های فرار پیشرفته آسیب‌پذیر می‌کند.
- **اکسپلویت‌های روز صفر:**
  - نقص‌های کشف‌نشده در الگوریتم‌ها یا پیکربندی‌های IDS فرصت‌هایی برای مهاجمان فراهم می‌کنند تا پایگاه‌هایی ایجاد کنند.
- **وابستگی بیش از حد به خودکارسازی:**
  - سیستم‌هایی که تنها به ابزارهای خودکار برای اعتبارسنجی یا طبقه‌بندی متکی هستند، ممکن است توسط ترافیک مصنوعی بسیار واقع‌گرایانه فریب بخورند.

۵-۳۷-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره شبکه هدف، از جمله معماری، الگوهای ترافیکی معمول و پیکربندی IDS استفاده می‌کند.
- **تحلیل و آموزش:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های ترافیکی تاریخی مستقر کرده و بر روی الگوهای رفتار قانونی آموزش می‌دهد تا استراتژی‌های فرار را بهبود بخشند.
- **توسعه استراتژی فرار:**
  - مهاجم الگوهای ترافیکی را طراحی کرده که با فعالیت عادی هماهنگ باشند و در عین حال اهداف مهاجم، مانند استخراج داده یا ارتباطات فرمان و کنترل (C2)، را پیش ببرند.
- **آزمایش و بهبود:**
  - مهاجم ترافیک طراحی‌شده را در محیط‌های شبیه‌سازی‌شده IDS آزمایش کرده تا اطمینان حاصل شود که از تشخیص فرار می‌کند و استراتژی‌ها را بر این اساس بهبود بخشید.
- **استقرار:**
  - مهاجم برنامه فرار ترافیکی را با ارسال payloadهای مخرب در حالی که بدون تشخیص باقی

می‌مانند، اجرا می‌کند.

- **پایداری:**
  - مهاجم الگوهای ترافیک را به‌طور مداوم بر اساس بازخوردهای لحظه‌ای به‌روزرسانی و تطبیق داده تا دسترسی بلندمدت حفظ شود.

۲-۳۷-۶ شانه‌های احتمال نفوذ

- **ناهنجاری‌های ظریف ترافیک:**
  - انحرافات جزئی در حجم ترافیک، زمان‌بندی یا ساختار که از هنجارهای مورد انتظار منحرف می‌شوند.
- **افزایش استفاده از منابع:**
  - مصرف بیشتر CPU، حافظه یا پهنای باند به دلیل اجرای اسکریپت‌های فرار خودکار در پس‌زمینه.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر یا سیستم ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.
- **ترافیک خروجی غیرمنتظره:**
  - انحراف در ترافیک خروجی، مانند اتصالات به آدرس‌های IP یا دامنه‌های ناشناخته، که ممکن است نشان‌دهنده استخراج داده یا ارتباطات فرمان و کنترل (C2) باشد.

۲-۳۷-۷ تأثیرات

- **نقض داده‌ها:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به ضررهای مالی یا آسیب شهرت می‌شود.
- **اختلال عملیاتی:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل اعتبارنامه‌های نفوذشده یا آسیب‌پذیری‌های سوءاستفاده‌شده.

- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلالات خدمات که باعث آسیب‌های مالی می‌شوند.
- **جریمه‌های نظارتی:**
  - عدم انطباق با مقررات حفاظت از داده‌ها، که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.
- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا فرم‌ورهای حیاتی، که سیستم‌ها را غیرقابل استفاده می‌کند.

۸-۳۷-۲ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - راه‌حل‌های IDS مبتنی بر هوش مصنوعی را مستقر کنید که بر روی مجموعه‌داده‌های بزرگی از ترافیک قانونی و مخرب آموزش دیده‌اند تا دقت را بهبود بخشند و مثبت‌های کاذب را کاهش دهند.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص انحرافات ظریف در الگوهای ترافیک که نشان‌دهنده تلاش‌های فرار هستند استفاده کنید.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا درباره تکنیک‌های نوظهور فرار و تهدیدات تطبیقی به‌روز بمانید.

- **به‌روزرسانی‌های منظم:**
  - تمام مجموعه‌قوانین IDS، فرم‌ورها و پیکربندی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده توسط ترافیک فرار اصلاح شوند.
- **رمزنگاری قوی:**
  - ارتباطات و داده‌های حساس را رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده کاهش یابد.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش دسترسی غیرمجاز محدود شود و

تأثیر فرار موفق کاهش یابد.

- **آموزش کاربران:**
  - کارمندان و کاربران را در مورد تشخیص علائم فعالیت مخرب آموزش دهید و فرهنگ هوشیاری را تقویت کنید.
- **لیست سیاه پیش‌گیرانه:**
  - از تحلیل‌های پیش‌بینانه برای سیاه‌کردن دامنه‌ها یا IPهای احتمالاً مخرب بر اساس الگوهای مشاهده‌شده استفاده کنید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت فرار موفق را تشخیص، مهار و کاهش دهید.
- **آزمایش تخاصمی:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های در حال تحول فرار استفاده کنید.

۲-۳۸ سازگارسازی حمله

سازگارسازی حملات[1] مبتنی بر هوش مصنوعی به استفاده از هوش مصنوعی برای تنظیم و بهینه‌سازی پویای حملات سایبری در پاسخ به تغییر شرایط، اقدامات دفاعی یا عوامل محیطی اشاره دارد. برخلاف حملات سنتی که به استراتژی‌های از پیش تعریف‌شده یا روش‌های ایستا متکی هستند، سازگاری حملات مبتنی بر هوش مصنوعی از الگوریتم‌های یادگیری ماشین و یادگیری عمیق استفاده می‌کند تا به‌طور مداوم از تعاملات با سیستم‌های هدف یاد بگیرند، تکنیک‌های حمله را بهبود بخشند و بر اقدامات ضد حمله غلبه کنند. این امر باعث می‌شود چنین حملاتی بسیار انعطاف‌پذیر، مقاوم و دفاع در برابر آن‌ها دشوار باشد.

سازگارسازی حملات مبتنی بر هوش مصنوعی نگران‌کننده است زیرا به مهاجمان امکان می‌دهد تا حتی مکانیزم‌های امنیتی پیشرفته مانند سیستم‌های تشخیص نفوذ (IDS)، فایروال‌ها و ابزارهای تحلیل رفتاری را با تکامل تاکتیک‌های خود در زمان واقعی دور بزنند.

۲-۳۸-۱ ویژگی‌های کلیدی

- **یادگیری پویا:**
  - مدل‌های هوش مصنوعی بازخورد از تلاش‌های قبلی را تحلیل می‌کنند تا حملات آینده را

1. attack adaptation

بهبود بخشند و با دفاع‌های جدید یا تغییرات در محیط هدف سازگار شوند.

- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند شناسایی آسیب‌پذیری‌ها، ایجاد payloadها و اجرای حملات را خودکار می‌کنند و تلاش دستی را کاهش می‌دهند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و رفتار خود را برای اجتناب از تشخیص یا سوءاستفاده از نقاط ضعف تازه‌یافته تغییر دهند.
- **پنهان‌کاری:**
  - با تقلید از فعالیت‌های قانونی یا ترکیب شدن با نویز پس‌زمینه، حملات مبتنی بر هوش مصنوعی برای مدت‌های طولانی کشف نشده باقی می‌مانند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین سیستم یا کاربر را به‌طور همزمان هدف‌گیری کنند، که این امر دامنه و تأثیر حمله را افزایش می‌دهد.

۲-۳۸-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌های داده‌ای از حملات موفق و ناموفق آموزش می‌دهد تا الگوها را شناسایی کند و تلاش‌های آینده را بهبود بخشد.
  - یادگیری بدون نظارت: ناهنجاری‌ها یا خوشه‌ها را در رفتار سیستم شناسایی می‌کند تا آسیب‌پذیری‌ها یا مکانیزم‌های دفاعی بالقوه را کشف کند.
  - یادگیری تقویتی: استراتژی‌های حمله را با پاداش دادن به نتایج موفقیت‌آمیز و جریمه کردن شکست‌ها بهینه می‌کند و امکان بهبود مداوم را فراهم می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در ترافیک شبکه، رفتار کاربر یا لاگ‌های سیستم را تحلیل می‌کنند تا تکنیک‌های حمله را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: داده‌های مصنوعی اما واقع‌گرایانه، مانند ترافیک شبکه جعلی یا کد مخرب ایجاد می‌کنند تا قابلیت‌های سازگاری را آزمایش و بهبود بخشند.

- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند و به مهاجمان امکان می‌دهد تا انحرافات را تشخیص دهند یا موجودیت‌های قانونی را جعل کنند.
- **تکنیک‌های :فازی:**
  - هوش مصنوعی را با روش‌های فازی ترکیب می‌کند تا سیستم‌ها را به‌طور سیستماتیک برای رفتارهای غیرمنتظره یا خرابی‌ها آزمایش کند و نقص‌های منطقی زیرین را آشکار کند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند پیام‌های خطا، فایل‌های لاگ یا الگوهای ارتباطی را تحلیل می‌کند تا زمینه‌های اضافی درباره سیستم هدف را استنباط کند.

۳-۳۸-۲ دارایی‌های هدف

- **سیستم‌های شبکه:**
  - شبکه‌های سازمانی، پلتفرم‌های ابری یا اکوسیستم‌های اینترنت اشیا (IoT) که دارای سرویس‌های در معرض خطر، فایروال‌های نادرست پیکربندی‌شده یا پروتکل‌های قدیمی هستند.
- **برنامه‌های نرم‌افزاری:**
  - برنامه‌های وب، برنامه‌های موبایل یا نرم‌افزارهای دسکتاپ که حاوی آسیب‌پذیری‌هایی مانند تزریق SQL، اسکریپت‌گذاری بین‌سایتی یا سرریز بافر هستند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، سیستم‌های تأمین آب، شبکه‌های حمل‌ونقل یا تأسیسات بهداشتی که حملات سازگار می‌توانند منجر به اختلالات قابل توجهی شوند.
- **سیستم‌های رمزنگاری:**
  - طرح‌های رمزنگاری سفارشی یا پروتکل‌هایی که در نرم‌افزار یا سخت‌افزار پیاده‌سازی شده‌اند و ممکن است نقص‌های کشف‌نشده داشته باشند.
- **مؤلفه‌های زنجیره تأمین:**
  - کتابخانه‌ها، وابستگی‌ها یا مؤلفه‌های متن‌باز شخص ثالث که در سیستم‌های بزرگتر ادغام شده‌اند و ممکن است آسیب‌پذیری‌ها را معرفی کنند.

۲-۳۸-۴ آسیب‌پذیری‌ها

- **دفاع‌های قدیمی:**
  - اقدامات امنیتی که تهدیدهای سازگار را در نظر نمی‌گیرند، ممکن است نتوانند به‌طور مؤثر تشخیص دهند یا پاسخ دهند.
- **پیکربندی ضعیف:**
  - فایروال‌ها، سرورها یا برنامه‌های نادرست پیکربندی‌شده که سرویس‌ها یا داده‌های حساس را در معرض خطر قرار می‌دهند.
- **نظارت ناکافی:**
  - عدم نظارت بلادرنگ یا ابزارهای تحلیل رفتاری، تشخیص حملات سازگار را دشوارتر می‌کند.
- **خطای انسانی:**
  - اشتباهات توسعه‌دهندگان، مدیران یا اپراتورها ممکن است به‌طور ناخواسته آسیب‌پذیری‌هایی را در سیستم‌ها ایجاد کنند.
- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - آسیب‌پذیری‌های کشف‌نشده که هنوز وصله‌ای برای آن‌ها ارائه نشده یا به‌طور عمومی افشا نشده‌اند، زمینه‌ای حاصلخیز برای حملات سازگار فراهم می‌کنند.

۲-۳۸-۵ سناریوی تهدید

- **شناسایی اولیه:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره سیستم هدف، از جمله معماری، پیکربندی و رفتار آن استفاده می‌کند.

۲-۳۸-۶ تحلیل و یادگیری:

- مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های جمع‌آوری‌شده مستقر می‌کند و الگوها یا ناهنجاری‌هایی را که نشان‌دهنده آسیب‌پذیری‌ها هستند، شناسایی می‌کند.
- **اجرای حمله:**
  - مهاجم یک حمله اولیه را بر اساس نقاط ضعف شناسایی‌شده راه‌اندازی می‌کند و به‌طور مداوم بازخورد درباره اثربخشی آن جمع‌آوری می‌کند.
- **سازگاری:**
  - مهاجم از یادگیری تقویتی یا سایر تکنیک‌های سازگار برای بهبود استراتژی حمله در زمان

واقعی استفاده می‌کند و بر موانع غلبه کرده و از دفاع‌ها اجتناب می‌کند.

- **پایداری:**
  - مهاجم درهای پشتی ایجاد می‌کند یا نقاط دسترسی را حفظ می‌کند تا امکان سوءاستفاده مکرر در طول زمان فراهم شود.
- **پوشاندن ردپا:**
  - مهاجم لاگ‌ها را تغییر می‌دهد، ردپاها را پاک می‌کند یا از تکنیک‌های دیگر برای اجتناب از تشخیص و حفظ پنهان‌کاری استفاده می‌کند.

۷-۳۸-۲ نشانه‌های احتمال نفوذ

- **الگوهای فعالیت غیرعادی:**
  - ناهنجاری‌ها در ترافیک شبکه، زمان‌های ورود یا استفاده از منابع که از رفتار پایه منحرف می‌شوند.
- **افزایش استفاده از منابع:**
  - مصرف بالاتر CPU یا حافظه به دلیل فعالیت‌های اسکن یا تحلیل خودکار.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول سیستم ناسازگار است، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به احراز هویت ناموفق، قفل شدن حساب‌ها یا فعالیت‌های غیرمنتظره.
- **تغییرات غیرمنتظره:**
  - تغییرات ناگهانی در پیکربندی سیستم، مجوزهای فایل یا تنظیمات شبکه که توسط کاربران قانونی آغاز نشده‌اند.

۸-۳۸-۲- تأثیرات

- **اختلال در عملیات:**
  - حملات سازگار می‌توانند با سوءاستفاده از آسیب‌پذیری‌های در حال تکامل، خدمات یا سیستم‌های حیاتی کسب‌وکار را مختل کنند.

- **نشت داده:**
  - سرقت داده‌های شخصی، مالی یا مالکیت فکری که منجر به خسارات مالی قابل توجه یا آسیب به اعتبار می‌شود.
- **خسارات مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلال در خدمات که منجر به آسیب مالی می‌شود.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و اقدامات قانونی می‌شود.
- **تشدید تهدیدات:**
  - حملات سازگار می‌توانند به اشکال پیچیده‌تری مانند حرکت جانبی، افزایش امتیاز یا حملات چندوجهی تبدیل شوند.

۹-۳۸-۲ راه‌های مقابله

- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص تغییرات ظریف در رفتار سیستم یا کاربر که نشان‌دهنده حملات سازگار هستند، استفاده کنید.
- **نظارت مداوم:**
  - راه‌حل‌های نظارت بلادرنگ را پیاده‌سازی کنید که قادر به تشخیص و پاسخ به تهدیدات در حال تکامل در محیط‌های پویا هستند.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرم‌ورها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده را برطرف کنید و از تهدیدات سازگار جلوتر بمانید.
- **هوشمندسازی تهدیدات پیشرفته:**
  - از پلتفرم‌های هوشمندسازی تهدیدات استفاده کنید تا درباره بردارهای حمله نوظهور و تکنیک‌های سازگار به‌روز بمانید.
- **آموزش متقابل:**
  - سیستم‌های امنیتی را با استفاده از یادگیری ماشین متقابل آموزش دهید تا سناریوهای حمله سازگار را شبیه‌سازی کرده و از آن‌ها دفاع کند.

- **دفاع چندلایه:**
  - ترکیبی از فایروال‌ها، سیستم‌های تشخیص/پیشگیری از نفوذ (IDS/IPS) و ابزارهای محافظت از نقاط پایانی را مستقر کنید تا دفاع‌های لایه‌ای ایجاد کنید.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد تشخیص علائم حملات سازگار و رعایت بهداشت امنیت سایبری آموزش دهید.
- **سیستم‌های پاسخ خودکار:**
  - سیستم‌های پاسخ به حوادث خودکار توسعه دهید که بتوانند به‌سرعت به تهدیدات تشخیص‌داده‌شده واکنش نشان دهند و پنجره فرصت را برای مهاجمان به حداقل برسانند.
- **تمرینات تیم قرمز:**
  - تمرینات منظم تیم قرمز را انجام دهید تا حملات سازگار را شبیه‌سازی کرده و مقاومت اقدامات دفاعی را آزمایش کنید.

۲-۳۹ سرقت رمز عبور

حمله سرقت رمز عبور[1] شامل تلاش مهاجم برای به‌دست آوردن مستقیم رمزهای عبور از یک سیستم، کاربر یا شبکه بدون نیاز به حدس زدن یا استفاده از روش‌های بروت فورس است. در حمله سرقت رمز عبور مبتنی بر هوش مصنوعی، از هوش مصنوعی برای افزایش کارایی و پنهان‌کاری تکنیک‌های سنتی مانند keylogging، فیشینگ، رونوشت کردن اعتبارنامه‌ها یا سوءاستفاده از آسیب‌پذیری‌های نرم‌افزار یا سخت‌افزار استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند شناسایی اهداف باارزش، بهینه‌سازی روش‌های استخراج داده و فرار از مکانیزم‌های تشخیص را خودکار کنند.

- حملات سرقت رمز عبور مبتنی بر هوش مصنوعی می‌توانند احراز هویت چندعاملی را دور بزنند، از روان‌شناسی انسان از طریق مهندسی اجتماعی سوءاستفاده کنند و به‌طور پویا با تغییرات در اقدامات امنیتی سازگار شوند.

۲-۳۹-۱ ویژگی‌های کلیدی

- **دقت هدف‌گیری:**
  - هوش مصنوعی حساب‌ها یا سیستم‌های باارزش با دفاع‌های ضعیف را شناسایی می‌کند تا

1. Password stealing attack

تأثیر حمله را به حداکثر برساند.

- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند شناسایی آسیب‌پذیری‌ها، استقرار بدافزار و استخراج اعتبارنامه‌ها را خودکار می‌کنند و تلاش دستی را کاهش می‌دهند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در اقدامات امنیتی، رفتار کاربر یا فناوری‌های دفاعی سازگار شوند.
- **پنهان‌کاری:**
  - با تقلید از فعالیت‌های قانونی و اجتناب از الگوهای مشکوک، هوش مصنوعی اطمینان می‌دهد که حمله برای مدت طولانی‌تری کشف نشود.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین سیستم یا کاربر را به‌طور همزمان هدف‌گیری کنند، که این امر دامنه و آسیب‌پذیری حمله را افزایش می‌دهد.

۲-۳۹-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌های داده‌ای از آسیب‌پذیری‌های شناخته‌شده، حملات موفق و اعتبارنامه‌های سرقت‌شده آموزش می‌دهد تا فرصت‌های جدید را پیش‌بینی کند.
  - یادگیری بدون نظارت: الگوهای رفتار کاربر یا لاگ‌های سیستم را شناسایی می‌کند تا آسیب‌پذیری‌ها یا ناهنجاری‌های پنهان را کشف کند.
  - یادگیری تقویتی: استراتژی‌های حمله را با پاداش دادن به سرقت موفقیت‌آمیز اعتبارنامه‌ها و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در تعاملات کاربر، ترافیک شبکه یا رفتار برنامه‌ها را تحلیل می‌کنند تا هدف‌گیری را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: ایمیل‌ها، وب‌سایت‌ها یا payloadهای بدافزار مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا کاربران و سیستم‌ها را فریب دهند.

- **پردازش زبان طبیعی:**
  - الگوهای زبانی در ایمیل‌های فیشینگ، پیام‌های چت یا پست‌های شبکه‌های اجتماعی را تحلیل می‌کند تا طعمه‌های متقاعدکننده برای حملات مهندسی اجتماعی ایجاد کند.
  - از پردازش زبان طبیعی برای استخراج اطلاعات حساس از منابع عمومی مانند پروفایل‌های شبکه‌های اجتماعی استفاده می‌کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر استفاده می‌کند تا مهاجمان بتوانند انحرافات را تشخیص دهند یا کاربران قانونی را جعل کنند.
- **بینایی کامپیوتر:**
  - در screen scraping یا تشخیص نوری کاراکترها (OCR) برای استخراج رمزهای عبور از رابط‌های بصری یا اسناد استفاده می‌شود.

۳-۳۹-۲ دارایی‌های هدف

- **حساب‌های کاربری:**
  - حساب‌های ایمیل، پروفایل‌های شبکه‌های اجتماعی، پلتفرم‌های بانکداری آنلاین و برنامه‌های سازمانی که اطلاعات حساس را ذخیره می‌کنند.
- **خدمات ابری:**
  - برنامه‌ها و APIهای میزبانی‌شده در ابر که در معرض اینترنت قرار دارند و رمزهای عبور ضعیف می‌توانند منجر به دسترسی غیرمجاز شوند.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - لوازم خانگی هوشمند، دستگاه‌های پوشیدنی و سنسورهای صنعتی که اغلب با رمزهای عبور پیش‌فرض یا ضعیف پیکربندی شده‌اند.
- **سیستم‌های سازمانی:**
  - شبکه‌های شرکتی، سرورهای فایل و سیستم‌های پایگاه‌داده که داده‌های حیاتی کسب‌وکار را ذخیره می‌کنند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، سیستم‌های تأمین آب و شبکه‌های حمل‌ونقل که برای عملکرد به احراز هویت امن متکی هستند.

۴-۳۹-۲ آسیب‌پذیری‌ها

- **مکانیزم‌های احراز هویت ضعیف:**
  - احراز هویت تک‌عاملی یا الگوریتم‌های تولید توکن قابل پیش‌بینی، سرقت اعتبارنامه‌ها را برای مهاجمان آسان‌تر می‌کند.
- **عدم محافظت از نقاط پایانی:**
  - عدم وجود ابزارهای قوی ضدویروس، ضدبدافزار یا تشخیص و پاسخ نقطه پایانی (EDR) به keyloggerها یا credential dumperها اجازه می‌دهد آزادانه عمل کنند.
- **خطای انسانی:**
  - کاربرانی که در دام کلاهبرداری‌های فیشینگ می‌افتند، روی لینک‌های مخرب کلیک می‌کنند یا فایل‌های آلوده دانلود می‌کنند، به‌طور ناخواسته اعتبارنامه‌های خود را در معرض خطر قرار می‌دهند.
- **نرم‌افزارهای قدیمی:**
  - استفاده از سیستم‌های عامل، برنامه‌ها یا ریزبرنامه‌های قدیمی که حاوی آسیب‌پذیری‌های شناخته‌شده قابل سوءاستفاده توسط بدافزار هستند.
- **نظارت ناکافی:**
  - روش‌های ضعیف ثبت لاگ یا عدم نظارت بلادرنگ، تشخیص و پاسخ به تلاش‌های سرقت اعتبارنامه را دشوار می‌کند.

۵-۳۹-۲ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره هدف، از جمله آدرس‌های ایمیل، نام‌های کاربری یا جزئیات شخصی عمومی استفاده می‌کند.
- **انتخاب بردار سوءاستفاده:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل محیط هدف و تعیین مؤثرترین بردار سوءاستفاده (مانند فیشینگ، بدافزار یا سوءاستفاده از آسیب‌پذیری) مستقر می‌کند.
- **استقرار بدافزار:**
  - مهاجم بدافزارهای سفارشی‌سازی‌شده مانند keyloggerها، را با استفاده از ایمیل‌های فیشینگ، دانلودهای مخرب یا کیت‌های سوءاستفاده تحویل می‌دهد.

- **استخراج اعتبارنامه:**
  - مهاجم از هوش مصنوعی برای خودکارسازی استخراج اعتبارنامه‌های ذخیره‌شده، مانند آن‌هایی که در مرورگرها، برنامه‌ها یا حافظه ذخیره شده‌اند، استفاده می‌کند.
- **خروج داده:**
  - اعتبارنامه‌های سرقت‌شده را به‌طور ایمن به سرور مهاجم منتقل می‌کند و از تشخیص توسط سیستم‌های تشخیص نفوذ (IDS) اجتناب می‌کند.
- **پایداری:**
  - درهای پشتی ایجاد می‌کند یا نقاط دسترسی را حفظ می‌کند تا دسترسی مکرر در طول زمان را تسهیل کند

۶-۳۹-۲ نشانه‌های احتمال نفوذ

- **ترافیک خروجی غیرمنتظره:**
  - ناهنجاری‌ها در ترافیک خروجی شبکه، مانند اتصالات به آدرس‌های IP یا دامنه‌های ناشناخته.
- **افزایش استفاده از CPU یا حافظه:**
  - مصرف منابع بیشتر از حد طبیعی به دلیل اجرای بدافزار در پس‌زمینه.
- **تلاش‌های ورود غیرمعمول:**
  - ورود از مکان‌ها، دستگاه‌ها یا زمان‌های ناآشنا.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر ناسازگار است، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ایمیل‌های فیشینگ:**
  - ایمیل‌های مشکوک حاوی پیوست‌ها یا لینک‌های مخرب که برای سرقت اعتبارنامه طراحی شده‌اند.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به احراز هویت ناموفق، قفل شدن حساب‌ها یا فعالیت‌های غیرمنتظره.

۲-۳۹-۷ تأثیرات

- **دسترسی غیرمجاز:**
  - سرقت موفقیت‌آمیز اعتبارنامه‌ها منجر به دسترسی غیرمجاز به حساب‌ها یا سیستم‌های حساس می‌شود و امکان سوءاستفاده بیشتر را فراهم می‌کند.
- **نشت داده:**
  - سرقت داده‌های شخصی، مالی یا مالکیت فکری که منجر به خسارات مالی قابل توجه یا آسیب به اعتبار می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل نفوذ به اعتبارنامه‌ها.
- **خسارات مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلال در خدمات که منجر به خسارات مالی می‌شود.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.

۲-۳۹-۸ راه‌های مقابله

- **احراز هویت چندعاملی:**
  - مراحل تأیید اضافی مانند کدهای یک‌بارمصرف یا اسکن‌های بیومتریک را فراتر از رمز عبور الزامی کنید تا از دسترسی غیرمجاز جلوگیری شود.
- **محافظت از نقاط پایانی:**
  - راه‌حل‌های پیشرفته ضدویروس، ضدبدافزار و تشخیص و پاسخ نقطه پایانی (EDR) را مستقر کنید تا فعالیت‌های مخرب را تشخیص و مسدود کنند.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، ریزبرنامه‌ها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده را برطرف کنید.
- **آموزش آگاهی امنیتی:**
  - کارکنان و کاربران را در مورد تشخیص تلاش‌های فیشینگ، اجتناب از دانلودهای مخرب و رعایت بهداشت امنیت سایبری آموزش دهید.

- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش بدافزار یا دسترسی غیرمجاز محدود شود.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای ورود غیرمعمول یا فعالیت‌های مشکوک که نشان‌دهنده سرقت اعتبارنامه هستند، استفاده کنید.
- **هش کردن و نمک زدن رمز عبور:**
  - رمزهای عبور را با استفاده از الگوریتم‌های هش قوی مانند Argon2، bcrypt یا PBKDF2 به‌طور ایمن ذخیره کنید و برای هر کاربر از نمک‌های منحصر به فرد استفاده کنید.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **CAPTCHA و تشخیص ربات:**
  - چالش‌های CAPTCHA یا سیستم‌های پیشرفته تشخیص ربات را پیاده‌سازی کنید تا از تلاش‌های خودکار سرقت اعتبارنامه جلوگیری شود.

۲-۴۰ سرقت اعتبارنامه

حمله سرقت اعتبارنامه[1] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای خودکارسازی و بهبود فرآیند سرقت اطلاعات ورود به سیستم، مانند نام‌های کاربری، رمزهای عبور، کلیدهای API یا توکن‌ها است. با استفاده از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند حجم عظیمی از داده‌ها از شبکه‌های اجتماعی، سوابق عمومی، نشت‌های داده و ترافیک شبکه را تحلیل کنند تا آسیب‌پذیری‌ها را شناسایی کنند، پیام‌های فیشینگ ایجاد کنند و مکانیزم‌های احراز هویت ضعیف را مورد سوءاستفاده قرار دهند. این نوع حمله به‌خصوص خطرناک است زیرا اعتماد به سیستم‌های احراز هویت را تضعیف می‌کند و به مهاجمان امکان می‌دهد تا به‌طور غیرمجاز به حساب‌ها، سیستم‌ها یا سرویس‌های حساس دسترسی پیدا کنند.

حملات سرقت اعتبارنامه مبتنی بر هوش مصنوعی به‌طور فزاینده‌ای پیچیده می‌شوند و دفاع‌های سنتی مانند احراز هویت چندعاملی، تحلیل رفتاری و سیستم‌های تشخیص نفوذ (IDS) را دور می‌زنند.

1. Credentials stealing

۲-۴۰-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند شناسایی اهداف، جمع‌آوری اطلاعات و اجرای سرقت اعتبارنامه را خودکار می‌کند، که تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **مقیاس‌پذیری:**

الگوریتم‌های یادگیری ماشین به مهاجمان امکان می‌دهند تا چندین کاربر یا سیستم را به‌طور همزمان هدف قرار دهند، که دامنه حمله را گسترش می‌دهد.

- **انطباق‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در اقدامات امنیتی، رفتار کاربران یا فناوری‌های دفاعی سازگار شوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا افراد یا سیستم‌های خاص با اعتبارهای با ارزش بالا را با دقت شناسایی کنند.
- **پنهان‌کاری:**
  - با تقلید از فعالیت‌های قانونی یا ترکیب شدن با نویز پس‌زمینه، حملات مبتنی بر هوش مصنوعی برای مدت‌های طولانی شناسایی نمی‌شوند.

۲-۴۰-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های برچسب‌دار از تلاش‌های موفق و ناموفق سرقت اعتبارنامه آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در رفتار کاربران یا لاگ‌های سیستم را شناسایی می‌کند تا نقاط ضعف احتمالی را آشکار کند.
  - یادگیری تقویتی: پارامترهای حمله را با پاداش دادن به نتایج موفق و جریمه کردن شکست‌ها بهینه می‌کند، که امکان بهبود مستمر را فراهم می‌کند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی مانند ایمیل‌ها، چت‌لاگ‌ها یا پست‌های شبکه‌های اجتماعی را تحلیل می‌کند تا بینش‌ها، احساسات یا اطلاعات شخصی (PII) را استخراج کند.

- پیام‌های فیشینگ بسیار شخصی‌سازی‌شده ایجاد می‌کند که سبک‌های ارتباطی مورد اعتماد را تقلید می‌کنند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در تعاملات کاربران، ترافیک شبکه یا رفتار برنامه‌ها را تحلیل می‌کنند تا فرصت‌های سرقت اعتبارنامه را شناسایی کنند.
  - شبکه‌های مولد تخاصمی: قالب‌های فیشینگ مصنوعی اما واقع‌گرایانه یا صفحات ورود جعلی ایجاد می‌کنند تا قربانیان را فریب دهند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربران استفاده می‌کند تا مهاجمان بتوانند به‌طور متقاعدکننده کاربران قانونی را تقلید کنند.
- **داده‌کاوی:**
  - از هوش مصنوعی برای استخراج داده‌های عمومی از شبکه‌های اجتماعی، فروم‌ها یا پایگاه‌داده‌های نشت‌شده استفاده می‌کند تا پروفایل‌های دقیقی از اهداف بالقوه ایجاد کند.

۳-۴۰-۲ دارایی‌های هدف

- **حساب‌های کاربری:**
  - حساب‌های ایمیل، پروفایل‌های شبکه‌های اجتماعی، پلتفرم‌های بانکداری آنلاین یا برنامه‌های سازمانی که اطلاعات حساس را ذخیره می‌کنند.
- **سیستم‌های سازمانی:**
  - شبکه‌های شرکتی، سرورهای فایل و سیستم‌های پایگاه‌داده که در آن‌ها اعتبارهای به‌خطرافتاده می‌توانند منجر به نشت داده یا حرکت جانبی شوند.
- **خدمات ابری:**
  - برنامه‌ها و APIهای میزبانی‌شده در ابر که در معرض اینترنت قرار دارند و در آن‌ها اعتبارهای سرقت‌شده می‌توانند دسترسی غیرمجاز را امکان‌پذیر کنند.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - لوازم خانگی هوشمند، دستگاه‌های پوشیدنی یا سنسورهای صنعتی که اغلب با اعتبارهای پیش‌فرض یا ضعیف پیکربندی شده‌اند.

- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، سیستم‌های تأمین آب، شبکه‌های حمل‌ونقل یا تأسیسات بهداشتی که در آن‌ها سرقت اعتبارنامه می‌تواند اختلالات قابل توجهی ایجاد کند.

۴-۴۰-۲ آسیب‌پذیری‌ها

- **مکانیزم‌های احراز هویت ضعیف:**
  - احراز هویت تک‌عاملی یا الگوریتم‌های تولید توکن قابل پیش‌بینی، سرقت اعتبارنامه را برای مهاجمان آسان‌تر می‌کند.
- **خطای انسانی:**
  - کاربرانی که اطلاعات شخصی خود را در شبکه‌های اجتماعی به‌اشتراک می‌گذارند یا در دام کلاهبرداری‌های فیشینگ می‌افتند، ناخواسته اعتبارهای خود را در معرض خطر قرار می‌دهند.
- **کنترل‌های حریم خصوصی ناکافی:**
  - فقدان رمزنگاری قوی، ناشناس‌سازی یا کنترل‌های دسترسی، داده‌های حساس را در معرض دسترسی غیرمجاز قرار می‌دهد.
- **سیستم‌های قدیمی:**
  - استفاده از نرم‌افزارها، فرمور یا پروتکل‌های قدیمی که حاوی آسیب‌پذیری‌های شناخته‌شده قابل سوءاستفاده توسط سارقان اعتبار هستند.
- **فقدان نظارت:**
  - روش‌های ضعیف ثبت رویدادها یا فقدان نظارت در زمان واقعی، تشخیص و پاسخ به تلاش‌های سرقت اعتبارنامه را دشوار می‌کند.

۵-۴۰-۲ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره هدف، از جمله حضور آنلاین، عادات ارتباطی یا فعالیت‌های اخیر استفاده می‌کند.
- **جمع‌آوری داده:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های عمومی، ارتباطات رهگیری‌شده یا پایگاه‌داده‌های نشت‌شده مستقر کرده تا پروفایل دقیقی از هدف ایجاد کنند.

- **کمپین‌های فیشینگ:**
  - مهاجم از تکنیک‌های پردازش زبان طبیعی و یادگیری عمیق برای ایجاد پیام‌های فیشینگ بسیار متقاعدکننده که متناسب با ترجیحات یا زمینه هدف هستند، استفاده می‌کند.
- **استخراج اعتبار:**
  - مهاجم بدافزارهای سفارشی، کی‌لاگرها یا ابزارهای استخراج اعتبار را از طریق ایمیل‌های فیشینگ، دانلودهای مخرب یا کیت‌های اکسپلویت مستقر می‌کند.
- **خروج داده‌ها:**
  - مهاجم اعتبارهای سرقت‌شده را به‌طور ایمن به سرور مهاجم انتقال داده و از تشخیص توسط سیستم‌های تشخیص نفوذ (IDS) جلوگیری می‌کند.
- **پایداری:**
  - مهاجم درهای پشتی ایجاد می‌کند یا نقاط دسترسی را حفظ کرده تا امکان دسترسی مکرر در طول زمان فراهم شود.

۶-۴۰-۲ نشانه‌های احتمال نفوذ

- **ترافیک خروجی غیرمنتظره:**
  - ناهنجاری‌ها در ترافیک خروجی شبکه، مانند اتصالات به آدرس‌های IP یا دامنه‌های ناشناخته.
- **افزایش استفاده از CPU/حافظه:**
  - مصرف منابع بیشتر از حد طبیعی به دلیل اجرای بدافزار در پس‌زمینه.
- **تلاش‌های ورود غیرمعمول:**
  - ورود به سیستم از مکان‌ها، دستگاه‌ها یا زمان‌های ناشناخته.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربران ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ایمیل‌های فیشینگ:**
  - ایمیل‌های مشکوک حاوی پیوست‌ها یا لینک‌های مخرب که برای سرقت اعتبارنامه طراحی شده‌اند.

- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به احراز هویت ناموفق، قفل شدن حساب‌ها یا فعالیت غیرمنتظره.

۷-۴۰-۲ تأثیرات

- **دسترسی غیرمجاز:**
  - سرقت موفقیت‌آمیز اعتبار منجر به دسترسی غیرمجاز به حساب‌ها یا سیستم‌های حساس می‌شود و امکان سوءاستفاده بیشتر را فراهم می‌کند.
- **نشت داده‌ها:**
  - سرقت داده‌های شخصی، مالی یا مالکیت فکری که منجر به ضررهای مالی قابل توجه یا آسیب‌های اعتباری می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به‌دلیل اعتبارهای به‌خطرافتاده یا آسیب‌پذیری‌های مورد سوءاستفاده.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلال در خدمات که باعث آسیب‌های مالی می‌شوند.
- **جریمه‌های نظارتی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، که منجر به جریمه‌ها و اقدامات قانونی می‌شود.

۸-۲۰-۲ راه‌های مقابله

- **احراز هویت چندعاملی:**
  - مراحل تأیید اضافی مانند کدهای یک‌بارمصرف یا اسکن‌های بیومتریک را علاوه بر رمزهای عبور الزامی کنید تا از دسترسی غیرمجاز جلوگیری شود.
- **رمزنگاری:**
  - داده‌های حساس را در هنگام ذخیره‌سازی و انتقال رمزنگاری کنید تا از رهگیری و سوءاستفاده جلوگیری شود.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرمور و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده را

برطرف کنید.

- **آموزش آگاهی امنیتی:**
  - کارمندان و کاربران را در مورد شناسایی تلاش‌های فیشینگ، اجتناب از دانلودهای مخرب و رعایت بهداشت سایبری آموزش دهید.
- **نظارت رفتاری:**
  - از تحلیل رفتاری مبتنی بر هوش مصنوعی استفاده کنید تا الگوهای ورود غیرمعمول یا فعالیت‌های مشکوک که نشان‌دهنده سرقت اعتبارنامه هستند، شناسایی شوند.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش دسترسی غیرمجاز یا جمع‌آوری داده محدود شود.
- **مدیریت اعتبار:**
  - سیاست‌های قوی رمز عبور را پیاده‌سازی کنید، استفاده مجدد از اعتبارها در سیستم‌های مختلف را ممنوع کنید و به‌روزرسانی‌های منظم را الزامی کنید.
- **CAPTCHA و تشخیص بات:**
  - چالش‌های CAPTCHA یا سیستم‌های پیشرفته تشخیص بات را مستقر کنید تا تلاش‌های خودکار سرقت اعتبارنامه را مسدود کنید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی‌ای توسعه دهید و پیاده‌سازی کنید تا سرقت اعتبارنامه را به‌سرعت شناسایی، مهار و کاهش دهید.
- **هوشمندسازی تهدیدات پیش‌گیرانه:**
  - از پلتفرم‌های هوشمند مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور سرقت اعتبارنامه و تهدیدات تطبیقی مطلع شوید.

۲-۴۱ ثبت کلید ضمنی

حمله ثبت کلید ضمنی[1] شامل استخراج اطلاعات حساس، مانند کلیدهای رمزنگاری یا رمزهای عبور، با تحلیل سیگنال‌های غیرمستقیم به جای رهگیری مستقیم داده‌ها است. برخلاف حملات ثبت کلید سنتی که به بدافزار برای ضبط کلیدهای فشرده شده متکی هستند، حمله ثبت کلید ضمنی مبتنی بر

1. Implicit Key Logging

هوش مصنوعی از هوش مصنوعی برای استنباط این اطلاعات از الگوهای رفتار سیستم، ترافیک شبکه یا تعاملات سخت‌افزاری استفاده می‌کند. با استفاده از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند سرنخ‌های ظریفی مانند زمان‌بندی، مصرف برق، انتشارات الکترومغناطیسی یا حتی نویز صوتی را تحلیل کنند تا کلیدهای مورد استفاده را استنباط کنند.

این نوع حمله می‌تواند بسیاری از اقدامات امنیتی سنتی، از جمله رمزنگاری، فایروال‌ها و نرم‌افزارهای ضدویروس را با تمرکز بر متاداده‌ها یا اطلاعات کانال جانبی دور بزند.

۲-۴۱-۱ ویژگی‌های کلیدی

- **استخراج غیرمستقیم داده:**
  - هوش مصنوعی سیگنال‌های ثانویه، مانند زمان‌بندی، مصرف برق یا انتشارات الکترومغناطیسی را تحلیل می‌کند تا کلیدهای حساس را بدون دسترسی مستقیم به داده‌ها استنباط کند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند شناسایی آسیب‌پذیری‌ها را خودکار می‌کنند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در سخت‌افزار، نرم‌افزار یا مکانیزم‌های دفاعی سازگار شوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا بیت‌ها یا بایت‌های خاصی از داده‌های حساس را با دقت بالا شناسایی کنند.
- **پنهان‌کاری:**
  - با استفاده از سیگنال‌های ظریفی که اغلب نادیده گرفته می‌شوند، حملات ثبت کلید ضمنی مبتنی بر هوش مصنوعی می‌توانند حتی در محیط‌های به‌خوبی ایمن‌سازی‌شده نیز کشف نشوند.

۲-۴۱-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را آموزش می‌دهد تا الگوهای داده‌های کانال جانبی را تشخیص

دهند و آن‌ها را به مقادیر محرمانه مربوطه (مانند کلیدهای رمزنگاری) نگاشت کنند.

- یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در داده‌های کانال جانبی را بدون برچسب‌گذاری قبلی شناسایی می‌کند.
- یادگیری تقویتی: استراتژی حمله را با پاداش دادن به بازیابی موفقیت‌آمیز کلید و جریمه کردن شکست‌ها بهینه می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در داده‌های سری زمانی، مانند ردیابی برق یا انتشارات الکترومغناطیسی را تحلیل می‌کنند.
  - شبکه‌های مولد تخاصمی: داده‌های کانال جانبی مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا برای اهداف آموزشی یا آزمایش استحکام اقدامات دفاعی استفاده شوند.
- **پردازش سیگنال:**
  - تکنیک‌هایی مانند تبدیل‌های فوریه، تحلیل موجک یا تحلیل مؤلفه‌های اصلی (PCA) را برای پیش‌پردازش و استخراج ویژگی‌های معنادار از داده‌های خام کانال جانبی اعمال می‌کند.
- **تحلیل سری‌های زمانی:**
  - از روش‌های آماری و یادگیری ماشین برای تحلیل وابستگی‌های زمانی در سیگنال‌های کانال جانبی استفاده می‌کند تا رفتارهای آینده را پیش‌بینی کند.

۳-۴۱-۲ دارایی‌های هدف

- **سیستم‌های رمزنگاری:**
  - ماژول‌های امنیتی سخت‌افزاری (HSM)، کارت‌های هوشمند و دستگاه‌های تعبیه‌شده که الگوریتم‌های رمزنگاری را پیاده‌سازی می‌کنند.
- **خدمات ابری:**
  - ماشین‌های مجازی (VM) و کانتینرهایی که منابع را در پلتفرم‌های چندمستأجری به اشتراک می‌گذارند، جایی که حملات زمان‌بندی یا مبتنی بر کش می‌توانند اطلاعات حساس را فاش کنند.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های IoT با منابع محدود که محافظت‌های کمی در برابر ثبت کلید ضمنی دارند.

- **سیستم‌های پرداخت:**
  - ترمینال‌های فروش، دستگاه‌های خواننده کارت و سیستم‌های پرداخت بدون تماس که در برابر تحلیل برق یا نشت الکترومغناطیسی آسیب‌پذیر هستند.
- **سیستم‌های دولتی و نظامی:**
  - دستگاه‌های ارتباطی امن و تجهیزات رمزنگاری مورد استفاده در عملیات دفاعی و اطلاعاتی.

۲-۴۱-۴ آسیب‌پذیری‌ها

- **نشت سیگنال‌های فیزیکی:**
  - عملیات رمزنگاری سیگنال‌های کانال جانبی قابل اندازه‌گیری، مانند مصرف برق، تشعشعات الکترومغناطیسی یا تغییرات زمان‌بندی منتشر می‌کنند.
- **اقدامات دفاعی ضعیف:**
  - تکنیک‌های ناکافی پوشش، کورسازی یا تزریق نویز، سیستم‌ها را در برابر ثبت کلید ضمنی آسیب‌پذیر می‌کند.
- **منابع اشتراکی:**
  - محیط‌های چندمستأجری، مانند پلتفرم‌های ابری، به مهاجمان امکان می‌دهند تا از کش‌ها، حافظه یا پردازنده‌های اشتراکی برای استنباط اطلاعات حساس سوءاستفاده کنند.
- **خطای انسانی:**
  - انتخاب‌های طراحی ضعیف، پیکربندی‌های نادرست یا تست ناکافی در طول توسعه ممکن است آسیب‌پذیری‌هایی را ایجاد کند که از طریق ثبت کلید ضمنی قابل سوءاستفاده باشند.
- **عدم نظارت:**
  - عدم نظارت مداوم بر فعالیت‌های غیرعادی کانال جانبی، به حملات امکان می‌دهد تا بدون تشخیص پیش‌بروند.

۲-۴۱-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از سنسورهای تخصصی یا ابزارهای ابزاردقیق برای ضبط سیگنال‌های کانال جانبی، مانند ردیابی برق، انتشارات الکترومغناطیسی یا اندازه‌گیری‌های زمان‌بندی استفاده می‌کند.
- **پیش‌پردازش:**
  - مهاجم تکنیک‌های پردازش سیگنال را برای پاک‌سازی و نرمال‌سازی داده‌های جمع‌آوری‌شده

اعمال می‌کند، نویز را حذف کرده و ویژگی‌های مرتبط را بهبود می‌بخشد.

- **آموزش مدل:**
  - مهاجم مدل‌های یادگیری ماشین را بر روی مجموعه‌های داده برچسب‌خورده آموزش می‌دهد تا سیگنال‌های کانال جانبی را با مقادیر محرمانه (مانند کلیدهای رمزنگاری) مرتبط کند.
- **اجرای حمله:**
  - مهاجم مدل آموزش‌دیده را برای تحلیل داده‌های کانال جانبی در زمان واقعی و استخراج اطلاعات حساس مستقر می‌کند.
- **پس‌پردازش:**
  - مهاجم داده‌های استخراج‌شده را با استفاده از تکنیک‌های پس‌پردازش بهبود می‌بخشد تا دقت را افزایش داده و خطاها را کاهش دهد.
- **سوءاستفاده:**
  - مهاجم از اسرار بازیابی‌شده برای اهداف مخرب، مانند رمزگشایی ارتباطات حساس یا جعل کاربران قانونی استفاده می‌کند.

۶-۴۱-۲ نشانه‌های احتمال نفوذ

- **فعالیت غیرعادی سنسور:**
  - وجود سنسورها یا ابزارهای نظارت غیرمجاز در نزدیکی سیستم‌های حیاتی.
- **مصرف برق غیرعادی:**
  - انحرافات در الگوهای مصرف برق که با عملیات یا رویدادهای خاص همبستگی دارند.
- **انتشارات الکترومغناطیسی غیرمنتظره:**
  - افزایش نویز الکترومغناطیسی یا تداخل در نزدیکی تجهیزات حساس.
- **ناهنجاری‌های زمان‌بندی:**
  - زمان‌های اجرای ناسازگار برای عملیات رمزنگاری یا سایر فرآیندهای بحرانی امنیتی.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول سیستم ناسازگار است، مانند تلاش‌های مکرر برای دسترسی به منابع یا عملکردهای خاص.

۲-۴۱-۷ تأثیرات

- **به‌خطر افتادن محرمانگی:**
  - افشای داده‌های حساس، مانند کلیدهای رمزنگاری، رمزهای عبور یا اطلاعات شخصی.
- **خسارات مالی:**
  - دسترسی غیرمجاز به سیستم‌های مالی یا سرقت دارایی‌های دیجیتال.
- **اختلال در عملیات:**
  - اختلال در ارتباطات امن یا زیرساخت‌های حیاتی به دلیل به‌خطر افتادن سیستم‌های رمزنگاری.
- **آسیب به اعتبار:**
  - از دست دادن اعتماد مشتری و ارزش برند پس از یک نفوذ.
- **جریمه‌های قانونی:**
  - عدم رعایت مقررات حفاظت از داده‌ها که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.
- **خطرات ایمنی:**
  - تهدید جان انسان‌ها در صورت به‌خطر افتادن دستگاه‌های پزشکی، سیستم‌های حمل‌ونقل یا کنترل‌های صنعتی.

۲-۴۱-۸ راه‌های مقابله

- **پوشش و کورسازی:**
  - تصادفی‌سازی را به عملیات رمزنگاری اضافه کنید تا سیگنال‌های کانال جانبی را مبهم کرده و تحلیل آن‌ها را دشوارتر کنید.
- **تزریق نویز:**
  - نویز کنترل‌شده را به سیستم معرفی کنید تا اندازه‌گیری‌های کانال جانبی را مختل کرده و مفید بودن آن‌ها را کاهش دهید.
- **بهبود طراحی سخت‌افزار:**
  - از قطعات سخت‌افزاری تخصصی، مانند تراشه‌های مقاوم در برابر تحلیل تفاضلی برق (DPA)، برای به حداقل رساندن نشت استفاده کنید.
- **روش‌های کدنویسی ایمن:**
  - الگوریتم‌های زمان‌ثابت را پیاده‌سازی کنید و از شاخه‌های شرطی که می‌توانند اطلاعات

زمان‌بندی را فاش کنند، اجتناب کنید.

- **حسابرسی‌های منظم:**
  - حسابرسی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌های بالقوه ثبت کلید ضمنی را شناسایی و برطرف کنید.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرعادی فعالیت کانال جانبی و ایجاد هشدار استفاده کنید.
- **ایزوله‌سازی منابع:**
  - اطمینان حاصل کنید که منابع در محیط‌های چندمستأجری به‌درستی ایزوله شده‌اند تا از حملات cross-VM یا cross-container جلوگیری شود.
- **اقدامات امنیتی فیزیکی:**
  - سیستم‌های حیاتی را با موانع فیزیکی، محافظ‌ها یا محفظه‌های مقاوم در برابر دستکاری محافظت کنید تا دسترسی به سیگنال‌های کانال جانبی محدود شود.

۲-۴۲ یافتن مسیر

حمله یافتن مسیر[1] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای شناسایی، تحلیل و بهره‌برداری از مسیرهای بهینه برای حرکت در شبکه‌ها، سیستم‌ها یا برنامه‌ها است. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند فرآیند نقشه‌برداری از محیط‌های پیچیده، کشف آسیب‌پذیری‌ها و تعیین کارآمدترین مسیرها برای حرکت جانبی، خروج داده یا ارتباطات فرمان و کنترل (C2) را خودکار کنند. این نوع حمله به‌خصوص خطرناک است زیرا به مهاجمان امکان می‌دهد با بهره‌برداری از نقاط ضعف در توپولوژی شبکه یا پیکربندی سیستم‌ها، از دفاع‌های سنتی مانند فایروال‌ها، سیستم‌های تشخیص نفوذ (IDS) یا کنترل‌های دسترسی عبور کنند.

حملات یافتن مسیر اغلب به‌عنوان بخشی از کمپین‌های بزرگ‌تر، مانند تهدیدات پیشرفته پایدار (APTs) یا عملیات باج‌افزاری، استفاده می‌شوند که در آن‌ها شناسایی کوتاه‌ترین یا پنهان‌ترین مسیر به دارایی‌های حیاتی برای موفقیت ضروری است.

۲-۴۲-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند اسکن شبکه‌ها، تحلیل الگوهای ترافیک و شناسایی مسیرهای قابل

1. Path Finding Attack

اجرا را خودکار می‌کند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.

- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی به‌طور پویا تکامل می‌یابند و با تغییرات در پیکربندی شبکه، اقدامات دفاعی یا شرایط محیطی سازگار می‌شوند.
- **پنهان‌کاری:**
  - با تقلید از ترافیک قانونی یا بهره‌برداری از نقاط کور در ابزارهای نظارتی، هوش مصنوعی اطمینان می‌دهد که فعالیت‌های یافتن مسیر کشف نشوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند مسیرهای خاص با اهداف باارزش را شناسایی کنند و خطر تشخیص را به حداقل برسانند در حالی که تأثیر را به حداکثر می‌رسانند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد چندین شبکه یا سیستم را به‌طور همزمان نقشه‌برداری و بهره‌برداری کنند و آسیب‌های بالقوه را افزایش دهند.

۲-۴۲-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های تلاش‌های موفق و ناموفق یافتن مسیر آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در ترافیک شبکه یا لاگ‌های سیستم را شناسایی می‌کند تا آسیب‌پذیری‌ها یا مسیرهای پنهان را آشکار کند.
  - یادگیری تقویتی: تاکتیک‌های یافتن مسیر را با پاداش‌دهی به عبور موفق و جریمه‌کردن شکست‌ها بهینه می‌کند و امکان بهبود مداوم را فراهم می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و شبکه‌های عصبی گراف، وابستگی‌های زمانی و روابط بین گره‌ها را تحلیل می‌کنند تا مسیرهای بهینه را تعیین کنند.
  - شبکه‌های مولد تخاصمی: الگوهای ترافیک مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های فرار را آزمایش و بهبود بخشند.

- **تحلیل گراف:**
  - از نظریه گراف و یادگیری ماشین برای مدل‌سازی توپولوژی شبکه‌ها، شناسایی گره‌های کلیدی و پیش‌بینی موثرترین مسیرهای حرکت استفاده می‌کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا فعالیت‌های مخرب را به‌طور نامحسوس با فعالیت‌های قانونی ترکیب کند.
- **تحلیل ترافیک:**
  - فراداده‌ها، زمان‌بندی یا اندازه بسته‌ها را تحلیل می‌کند تا اطلاعات حساس درباره ساختار شبکه‌ها یا الگوهای ارتباطی را استنباط کند.

۳-۴۲-۲ دارایی‌های هدف

- **شبکه‌های سازمانی:**
  - اینترانت‌های شرکتی که در آن‌ها یافتن مسیر می‌تواند منجر به دسترسی غیرمجاز به داده‌های حساس یا سیستم‌های حیاتی شود.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تأسیسات بهداشتی، مؤسسات مالی یا شبکه‌های حمل‌ونقل که در آن‌ها مسیرهای به‌خطرافتاده می‌توانند آسیب‌های قابل توجهی ایجاد کنند.
- **خدمات ابری:**
  - پلتفرم‌های ابری که برنامه‌ها، ذخیره‌سازی یا منابع محاسباتی را میزبانی می‌کنند و از طریق اعتبارنامه‌های به‌خطرافتاده یا پیکربندی‌های نادرست قابل دسترسی هستند.
- **شبکه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های IoT با منابع محدود که اغلب به‌عنوان پل‌هایی برای نفوذ بیشتر به اکوسیستم‌های بزرگ‌تر استفاده می‌شوند.
- **سیستم‌های زنجیره تأمین:**
  - فروشندگان، شرکا یا تأمین‌کنندگان شخص ثالث که حساب‌های به‌خطرافتاده آن‌ها می‌توانند به‌عنوان نقاط ورود برای حملات یافتن مسیر عمل کنند.

۳-۴۲-۲ آسیب‌پذیری‌ها

- **تقسیم‌بندی ضعیف شبکه:**
  - عدم تقسیم‌بندی قوی شبکه، سیستم‌ها را در برابر حرکت جانبی و عبور غیرمجاز آسیب‌پذیر می‌کند.
- **نظارت ناکافی:**
  - روش‌های ضعیف ثبت لاگ یا نظارت ناکافی در زمان واقعی، تشخیص فعالیت‌های یافتن مسیر را دشوارتر می‌کند.
- **خطای انسانی:**
  - پیکربندی نادرست فایروال‌ها، اعتبارنامه‌های مشترک یا آموزش ناکافی به‌طور ناخواسته آسیب‌پذیری‌هایی را ایجاد می‌کند که در طول یافتن مسیر قابل بهره‌برداری هستند.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های قدیمی یا نرم‌افزارهای منسوخ، سازمان‌ها را در برابر تکنیک‌های انطباق‌پذیر یافتن مسیر آسیب‌پذیر می‌کند.
- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - آسیب‌پذیری‌های کشف‌نشده در نرم‌افزار یا سخت‌افزار، فرصت‌هایی را برای مهاجمان فراهم می‌کنند تا پایگاه‌هایی ایجاد کنند و مسیرها را نقشه‌برداری کنند.

۴-۴۲-۲ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره شبکه هدف، از جمله معماری، الگوهای ترافیک معمول و پیکربندی IDS استفاده می‌کند.
- **نقشه‌برداری توپولوژی:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل توپولوژی شبکه به‌کار می‌گیرد تا گره‌های کلیدی، کانال‌های ارتباطی و آسیب‌پذیری‌های بالقوه را شناسایی کند.
- **بهینه‌سازی مسیر:**
  - مهاجم از یادگیری تقویتی یا تحلیل گراف برای تعیین کارآمدترین یا پنهان‌ترین مسیرها برای حرکت، خروج داده یا ارتباطات فرمان و کنترل (C2) استفاده می‌کند.

- **بهره‌برداری:**
  - مهاجم استراتژی یافتن مسیر برنامه‌ریزی‌شده را اجرا می‌کند و اطمینان می‌دهد که اقدامات با زمینه قربانی همسو بوده و خطر تشخیص را به حداقل می‌رساند.
- **پایداری:**
  - مهاجم رویکرد یافتن مسیر را بر اساس بازخورد به‌طور مداوم بهبود می‌بخشد و دسترسی بلندمدت یا تأثیر بر محیط‌های هدف را حفظ می‌کند.

۲-۴۲-۵ نشانه‌های احتمال نفوذ

- **الگوهای ترافیک غیرعادی:**
  - ناهنجاری‌ها در اندازه بسته‌ها، زمان‌های بین ورود یا جهت جریان‌ها که از رفتار پایه منحرف می‌شوند.
- **افزایش تاخیر:**
  - افزایش ناگهانی تاخیر در ارتباطات که ممکن است نشان‌دهنده رهگیری یا دستکاری در مسیرهای نقشه‌برداری‌شده باشد.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار عادی کاربر، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرعادی.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.
- **ترافیک خروجی غیرمنتظره:**
  - انحراف در ترافیک خروجی، مانند اتصال به آدرس‌های IP یا دامنه‌های ناشناخته، که ممکن است نشان‌دهنده ارتباطات فرمان و کنترل (C2) یا خروج داده باشد.

۲-۴۲-۶ تأثیرات

- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل ارتباطات مخدوش یا تزریق بدافزار.
- **نقض داده:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به

ضرر مالی یا آسیب به اعتبار می‌شود.

- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، کلاهبرداری یا اختلال در خدمات که باعث آسیب مالی می‌شوند.
- **آسیب به اعتبار:**
  - قربانیان حملات یافتن مسیر مبتنی بر هوش مصنوعی ممکن است به دلیل داده‌های افشا یا دستکاری‌شده، آسیب‌های اعتباری ببینند.
- **فساد سیستم:**
  - تزریق payloadهای مخرب به ارتباطات رهگیری‌شده می‌تواند پایگاه‌های داده، برنامه‌ها یا فریم‌ور را فاسد کند.

۲-۴۲-۷ راه‌های مقابله

- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا حرکت جانبی محدود شده و تأثیر یافتن مسیر موفق کاهش یابد.
- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های IDS/IPS مبتنی بر هوش مصنوعی را مستقر کنید که قادر به تشخیص الگوهای ترافیک غیرعادی نشان‌دهنده تلاش‌های یافتن مسیر هستند.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای نظارت بر انحرافات در فعالیت کاربر یا تعاملات سیستم استفاده کنید که ممکن است نشان‌دهنده قصد مخرب باشد.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فریم‌ورها و پیکربندی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده در یافتن مسیر اصلاح شوند.
- **رعایت اصول رمزنگاری:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حال انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده به حداقل برسد.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد شناسایی علائم خطر، مانند فعالیت‌های غیرعادی سیستم،

آموزش دهید و فرهنگ هوشیاری را تقویت کنید.

- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور یافتن مسیر و تهدیدات انطباق‌پذیر مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حملات یافتن مسیر را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.
- **دفاع پیشگیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های در حال تکامل یافتن مسیر استفاده کنید.
- **کنترل‌های دسترسی:**
  - کنترل‌های دسترسی سخت‌گیرانه، احراز هویت چندعاملی و اصل حداقل امتیاز را اعمال کنید تا دسترسی غیرمجاز محدود شود.

۲-۴۳ حرکت جانبی

حرکت جانبی[1] به تکنیکی اشاره دارد که مهاجمان از آن برای حرکت از یک سیستم اولیه‌ای که به آن نفوذ کرده‌اند به سایر سیستم‌های موجود در شبکه استفاده می‌کنند تا دسترسی و کنترل خود را گسترش دهند. در حمله حرکت جانبی مبتنی بر هوش مصنوعی، از هوش مصنوعی برای خودکارسازی و بهبود این فرآیند استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند توپولوژی شبکه را تحلیل کنند، اهداف باارزش را شناسایی کنند و تاکتیک‌های خود را به‌طور لحظه‌ای تطبیق دهند تا از تشخیص فرار کنند و تأثیر خود را به حداکثر برسانند.

این نوع حمله به‌ویژه خطرناک است زیرا به مهاجمان امکان می‌دهد به‌طور پنهانی در شبکه‌ها حرکت کنند، آسیب‌پذیری‌ها را به‌طور مؤثر سوءاستفاده کنند و حضور خود را در طول زمان حفظ کنند. حرکت جانبی مبتنی بر هوش مصنوعی می‌تواند از دفاع‌های سنتی مانند سیستم‌های تشخیص نفوذ (IDS)، فایروال‌ها یا ابزارهای تحلیل رفتاری عبور کند، که این امر مهار تهدید را برای مدافعان دشوارتر می‌سازد.

1. Lateral Movement

۱-۴۳-۲ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند شناسایی سیستم‌های آسیب‌پذیر، سوءاستفاده از نقاط ضعف و افزایش امتیازات را خودکار می‌کند، که باعث کاهش تلاش دستی و افزایش کارایی می‌شود.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی به‌طور پویا تکامل می‌یابند و با تغییرات در پیکربندی شبکه، اقدامات دفاعی یا شرایط محیطی سازگار می‌شوند.
- **پنهان‌کاری:**
  - با تقلید از رفتار کاربران قانونی یا رمزنگاری payloadها تا زمان اجرا، هوش مصنوعی اطمینان می‌دهد که حرکت جانبی بدون تشخیص باقی بماند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند سیستم‌ها یا حساب‌های خاصی را که دارای داده‌های باارزش یا امتیازات بالا هستند، شناسایی کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد به‌طور همزمان چندین سیستم یا شبکه را طی کنند، که آسیب بالقوه را افزایش می‌دهد.

۲-۴۳-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های موفق و ناموفق تلاش‌های حرکت جانبی آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا ناهنجاری‌ها را در ترافیک شبکه یا لاگ‌های سیستم شناسایی می‌کند تا آسیب‌پذیری‌های بالقوه را آشکار کند.
  - یادگیری تقویتی: تاکتیک‌های حرکت جانبی را با پاداش دادن به نتایج موفق و جریمه کردن شکست‌ها بهینه می‌کند، که امکان بهبود مستمر را فراهم می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و شبکه‌های عصبی گرافی، روابط پیچیده بین سیستم‌ها، کاربران و فرآیندها را تحلیل می‌کنند تا حرکت را هدایت کنند.

- شبکه‌های مولد تخاصمی: فعالیت‌های شبکه‌ای مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های فرار را آزمایش و بهبود بخشند.

- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار طبیعی کاربر یا سیستم استفاده می‌کند تا مهاجمان بتوانند فعالیت‌های مخرب را با فعالیت‌های قانونی ترکیب کنند.
- **تحلیل شبکه:**
  - از هوش مصنوعی برای نقشه‌برداری از توپولوژی شبکه، شناسایی دارایی‌های حیاتی و تعیین مسیرهای بهینه برای حرکت جانبی استفاده می‌کند.
- **داده‌کاوی:**
  - حجم عظیمی از داده‌ها را از سیستم‌های نفوذشده تحلیل می‌کند تا اعتبارنامه‌ها، آسیب‌پذیری‌ها یا روابط مورد اعتماد را کشف کند که حرکت را تسهیل می‌کنند.

۳-۴۳-۲ دارایی‌های هدف

- **شبکه‌های سازمانی:**
  - اینترانت‌های شرکتی که حرکت جانبی می‌تواند منجر به دسترسی غیرمجاز به داده‌های حساس یا سیستم‌های حیاتی شود.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، مراکز بهداشتی، مؤسسات مالی یا شبکه‌های حمل‌ونقل که حرکت می‌تواند آسیب‌های جدی ایجاد کند.
- **محیط‌های ابری:**
  - برنامه‌ها، ذخیره‌سازی یا منابع محاسباتی میزبانی‌شده در ابر که از طریق اعتبارنامه‌های نفوذشده یا پیکربندی‌های نادرست قابل دسترسی هستند.
- **شبکه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های اینترنت اشیا با منابع محدود که اغلب به عنوان پله‌هایی برای نفوذ بیشتر به اکوسیستم‌های بزرگتر استفاده می‌شوند.
- **سیستم‌های زنجیره تأمین:**
  - تأمین‌کنندگان، شرکا یا فروشندگان شخص ثالث که حساب‌های آن‌ها می‌تواند به عنوان نقاط ورودی برای حرکت جانبی مورد سوءاستفاده قرار گیرد.

۴-۴۳-۲ آسیب‌پذیری‌ها

- **مدیریت ضعیف اعتبارنامه‌ها:**
  - استفاده مجدد از رمزهای عبور، مکانیزم‌های احراز هویت ضعیف یا اعتبارنامه‌های ایمن‌سازی‌نشده فرصت‌هایی برای حرکت جانبی فراهم می‌کنند.
- **نظارت ناکافی:**
  - نبود نظارت لحظه‌ای یا ابزارهای تحلیل رفتاری، تشخیص علائم ظریف حرکت را دشوارتر می‌کند.
- **خطای انسانی:**
  - فایروال‌های نادرست پیکربندی‌شده، اعتبارنامه‌های مشترک یا آموزش ناکافی به‌طور ناخواسته آسیب‌پذیری‌هایی را در معرض سوءاستفاده در طول حرکت جانبی قرار می‌دهند.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های قدیمی یا نرم‌افزارهای منسوخ، سازمان‌ها را در برابر تکنیک‌های حرکت پیشرفته آسیب‌پذیر می‌کند.
- **سوءاستفاده از روز صفر:**
  - آسیب‌پذیری‌های کشف‌نشده در نرم‌افزار یا سخت‌افزار، پایگاه‌هایی برای مهاجمان فراهم می‌کنند تا حضور خود را ایجاد و گسترش دهند.

۵-۴۳-۲ سناریوی تهدید

- **نفوذ اولیه:**
  - مهاجم از طریق فیشینگ، بدافزار یا سایر بردارها به یک سیستم سطح پایین دسترسی پیدا کرده و پایگاهی در شبکه هدف ایجاد می‌کند.
- **شناسایی شبکه:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای نقشه‌برداری از توپولوژی شبکه، شناسایی دارایی‌های حیاتی و یافتن حساب‌ها یا سیستم‌های دارای امتیاز استفاده می‌کند.
- **جمع‌آوری اعتبارنامه:**
  - مهاجم الگوریتم‌های یادگیری ماشین را برای استخراج اعتبارنامه‌ها، توکن‌ها یا گواهی‌ها از سیستم‌های نفوذشده مستقر می‌کند.

- **افزایش امتیازات:**
  - مهاجم از اعتبارنامه‌های جمع‌آوری‌شده یا سوءاستفاده از آسیب‌پذیری‌ها برای دسترسی به سطوح بالاتر در شبکه استفاده می‌کند.
- **بهینه‌سازی مسیر:**
  - مهاجم از یادگیری تقویتی برای تعیین کارآمدترین و کم‌ترین مسیرهای قابل تشخیص برای حرکت جانبی استفاده می‌کند.
- **اجرا:**
  - مهاجم در شبکه به‌صورت جانبی حرکت کرده و به سیستم‌ها یا داده‌های اضافی دسترسی پیدا می‌کند و در عین حال از تشخیص فرار می‌کند.
- **پایداری:**
  - مهاجم درهای پشتی ایجاد کرده یا نقاط دسترسی را حفظ می‌کند تا حرکت مکرر و حضور بلندمدت تسهیل شود.

۶-۴۳-۲ شانه‌های احتمال نفوذ

- **الگوهای دسترسی غیرمعمول:**
  - ناهنجاری‌ها در زمان‌های ورود به سیستم، مکان‌ها یا استفاده از دستگاه که با کاربران قانونی مرتبط نیستند.
- **افزایش تلاش‌های احراز هویت:**
  - افزایش ناگهانی در تلاش‌های ورود ناموفق که با دسترسی موفق دنبال می‌شود، می‌تواند نشان‌دهنده جمع‌آوری اعتبارنامه یا حرکت جانبی باشد.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.
- **ترافیک شبکه غیرمنتظره:**
  - انحراف در ترافیک داخلی شبکه، مانند اتصالات بین سیستم‌هایی که معمولاً با هم ارتباط برقرار نمی‌کنند.

۲-۴۳-۷ تأثیرات

- **نقض داده‌ها:**
  - دسترسی غیرمجاز به اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به ضررهای مالی یا آسیب اعتباری می‌شود.
- **اختلال عملیاتی:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل اعتبارنامه‌های نفوذشده یا آسیب‌پذیری‌های سوءاستفاده‌شده.
- **ضررهای مالی:**

تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلالات خدمات که باعث آسیب‌های مالی می‌شوند.

- **جریمه‌های نظارتی:**

عدم انطباق با مقررات حفاظت از داده‌ها، که منجر به جریمه‌ها و اقدامات قانونی می‌شود.

- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا فرم‌ورهای حیاتی، که سیستم‌ها را غیرقابل استفاده می‌کند.

۲-۴۳-۸ راه‌های مقابله

- **احراز هویت چندعاملی:**
  - مراحل تأیید اضافی فراتر از رمزهای عبور، مانند کدهای یک‌بار مصرف یا اسکن‌های بیومتریک، را برای محافظت در برابر دسترسی غیرمجاز الزامی کنید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش دسترسی غیرمجاز محدود شود و تأثیر حرکت جانبی موفق کاهش یابد.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص الگوهای غیرمعمول فعالیت که نشان‌دهنده حرکت جانبی هستند استفاده کنید، حتی اگر اقدامات قانونی به نظر برسند.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرم‌ورها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده

مورد سوءاستفاده در طول حرکت جانبی اصلاح شوند.

- **مدیریت اعتبارنامه:**
  - سیاست‌های قوی رمز عبور را پیاده‌سازی کنید، استفاده مجدد از اعتبارنامه‌ها در سیستم‌های مختلف را ممنوع کنید و به‌روزرسانی‌های منظم را اجباری کنید.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا درباره تکنیک‌های نوظهور حرکت جانبی و تهدیدات تطبیقی به‌روز بمانید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حرکت جانبی را تشخیص، مهار و کاهش دهید.
- **تشخیص و پاسخ نقطه پایانی (EDR):**
  - راه‌حل‌های EDR را مستقر کنید که قادر به تشخیص و پاسخ به فعالیت‌های مشکوک در سطح نقطه پایانی هستند.
- **دفاع پیش‌گیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های در حال تحول حرکت جانبی استفاده کنید.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد تشخیص علائم نفوذ، مانند رفتار غیرمعمول سیستم یا درخواست‌های غیرمنتظره آموزش دهید.

۲-۴۴ استخراج داده‌ها

استخراج یا خروج داده‌ها[1] به انتقال غیرمجاز اطلاعات حساس از یک سیستم یا شبکه هدف به یک مکان خارجی که توسط مهاجم کنترل می‌شود، اشاره دارد. در یک حمله خروج داده‌ها مبتنی بر هوش مصنوعی، از هوش مصنوعی برای خودکارسازی و بهبود فرآیند شناسایی، استخراج و انتقال داده‌های ارزشمند در حالی که از تشخیص جلوگیری می‌شود، استفاده می‌شود. با استفاده از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند حجم عظیمی از داده‌ها را تحلیل کنند، اهداف با ارزش بالا را شناسایی کنند، تکنیک‌های خروج داده‌ها را بهینه کنند و با اقدامات دفاعی مانند تشخیص ناهنجاری،

1. Data exfiltration

رمزنگاری یا نظارت شبکه سازگار شوند.

این نوع حمله به‌خصوص خطرناک است زیرا می‌تواند دفاع‌های سنتی را دور بزند، به‌طور کارآمد در چندین سیستم مقیاس‌پذیر باشد و به‌طور پویا با محیط‌های در حال تغییر سازگار شود، که به مهاجمان امکان می‌دهد حجم زیادی از داده‌های حساس را بدون اینکه شناسایی شوند، سرقت کنند.

۱-۴۴-۲ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند شناسایی داده‌های حساس، ایجاد روش‌های خروج داده‌ها و اجرای انتقال‌ها را خودکار می‌کند، که تلاش دستی را کاهش می‌دهد.
- **مقیاس‌پذیری:**
  - الگوریتم‌های یادگیری ماشین به مهاجمان امکان می‌دهند تا چندین سیستم یا مجموعه داده‌ها را به‌طور همزمان هدف قرار دهند، که دامنه حمله را افزایش می‌دهد.
- **انطباق‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در اقدامات امنیتی، پیکربندی شبکه یا شرایط محیطی سازگار شوند.
- **پنهان‌کاری:**
  - با تقلید از الگوهای ترافیک قانونی یا رمزنگاری payloadها تا زمان انتقال، هوش مصنوعی اطمینان می‌دهد که فعالیت‌های مخرب در طول تحلیل شناسایی نشوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا انواع خاصی از داده‌های حساس، مانند اطلاعات شخصی، سوابق مالی یا مالکیت فکری را با دقت شناسایی کنند.

۲-۴۴-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های برچسب‌دار از انواع شناخته‌شده داده‌های حساس آموزش می‌دهد تا اهداف جدید برای خروج داده‌ها را شناسایی کند.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در ذخیره‌سازی داده‌ها، الگوهای دسترسی یا ترافیک شبکه را شناسایی می‌کند تا آسیب‌پذیری‌های پنهان را آشکار کند.
  - یادگیری تقویتی: استراتژی‌های خروج داده‌ها را با پاداش دادن به استخراج موفق داده و

جریمه کردن تشخیص یا شکست بهینه می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در ساختار داده‌ها، رفتار کاربران یا فعالیت شبکه را تحلیل می‌کنند تا تکنیک‌های خروج داده‌ها را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: جریان‌های داده مصنوعی اما واقع‌گرایانه یا الگوهای ارتباطی ایجاد می‌کنند تا قابلیت‌های دور زدن را تست و بهبود بخشند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی مانند اسناد، ایمیل‌ها یا چت‌لاگ‌ها را تحلیل می‌کند تا اطلاعات حساس مانند اعتبارها، اسرار تجاری یا ارتباطات محرمانه را استخراج کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا فعالیت‌های مخرب را با فعالیت‌های قانونی ترکیب کند.
- **داده‌کاوی:**
  - از هوش مصنوعی برای استخراج داده‌های عمومی از شبکه‌های اجتماعی، فروم‌ها یا پایگاه‌داده‌های نشت‌شده استفاده می‌کند تا پروفایل‌های دقیقی از اهداف بالقوه ایجاد کند یا نقاط ضعف در امنیت داده‌ها را شناسایی کند.

۳-۴۴-۲ دارایی‌های هدف

- **پایگاه‌داده‌های حساس:**
  - پایگاه‌داده‌های شرکتی که اطلاعات مشتریان، سوابق مالی یا مالکیت فکری را ذخیره می‌کنند.
- **ذخیره‌سازی ابری:**
  - خدمات میزبانی‌شده در ابر که داده‌های حساس در آن‌ها ذخیره شده و از طریق APIها یا رابط‌های وب قابل دسترسی هستند.
- **شبکه‌های سازمانی:**
  - شبکه‌های شرکتی که داده‌های حیاتی کسب‌وکار، از جمله سوابق کارمندان، برنامه‌های استراتژیک یا جزئیات عملیاتی را در خود جای داده‌اند.

- **دستگاه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های IoT با منابع محدود که داده‌های حساس را جمع‌آوری و انتقال می‌دهند و اغلب دارای کنترل‌های امنیتی محدودی هستند.
- **زیرساخت‌های حیاتی:**
  - سیستم‌هایی که شبکه‌های برق، تأسیسات بهداشتی، شبکه‌های حمل‌ونقل یا سایر خدمات ضروری را کنترل می‌کنند و در آن‌ها سرقت داده‌ها می‌تواند آسیب‌های قابل توجهی ایجاد کند.

۲-۴۴-۴ آسیب‌پذیری‌ها

- **رمزنگاری ضعیف:**
  - استفاده از پروتکل‌های رمزنگاری قدیمی یا نادرست پیاده‌سازی‌شده، داده‌های حساس را در طول انتقال در معرض خطر قرار می‌دهد.
- **نظارت ناکافی:**
  - فقدان نظارت در زمان واقعی یا ابزارهای تحلیل رفتاری، تشخیص الگوهای غیرمعمول دسترسی یا انتقال داده را دشوار می‌کند.
- **خطای انسانی:**
  - فایروال‌های نادرست پیکربندی‌شده، اعتبارهای مشترک یا آموزش ناکافی به‌طور ناخواسته آسیب‌پذیری‌هایی را ایجاد می‌کنند که توسط خروج داده‌ها مبتنی بر هوش مصنوعی مورد سوءاستفاده قرار می‌گیرند.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های قدیمی یا نرم‌افزارهای منسوخ، سازمان‌ها را در برابر تکنیک‌های پیشرفته خروج داده‌ها آسیب‌پذیر می‌کند.
- **سوءاستفاده از آسیب‌پذیری‌های روز صفر:**
  - آسیب‌پذیری‌های کشف‌نشده در نرم‌افزار یا سخت‌افزار، نقاط ورودی برای مهاجمان فراهم می‌کنند تا پایگاه‌هایی برای خروج داده‌ها ایجاد کنند.

۲-۴۴-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره سیستم هدف،

از جمله معماری آن، مکان‌های ذخیره‌سازی داده و الگوهای ترافیک معمول استفاده می‌کند.

- **شناسایی داده:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل مجموعه‌داده‌ها و شناسایی اهداف با ارزش بالا، مانند PII، داده‌های مالی یا اطلاعات اختصاصی مستقر می‌کند.
- **برنامه‌ریزی خروج داده‌ها:**
  - مهاجم از هوش مصنوعی برای طراحی روش‌های مخفیانه خروج داده‌ها، مانند رمزنگاری payloadها، تقسیم داده‌ها یا پنهان کردن انتقال‌ها به‌عنوان ترافیک قانونی استفاده می‌کند.
- **تست و بهبود:**
  - مهاجم تکنیک‌های برنامه‌ریزی‌شده خروج داده‌ها را در محیط‌های شبیه‌سازی‌شده تست کرده تا اطمینان حاصل شود که از تشخیص فرار می‌کنند و استراتژی‌ها را بر این اساس بهبود می‌بخشد.
- **استقرار:**
  - مهاجم برنامه خروج داده‌ها را با انتقال داده‌های حساس به سرورها یا مکان‌های ذخیره‌سازی خارجی در حالی که شناسایی نمی‌شوند، اجرا می‌کند.
- **پایداری:**
  - مهاجم دسترسی بلندمدت به سیستم‌ها یا حساب‌های به‌خطرافتاده را با تطبیق مستمر تکنیک‌های خروج داده‌ها برای فرار از دفاع‌های به‌روزرسانی‌شده حفظ می‌کند.

۶-۴۴-۲ نشانه‌های احتمال نفوذ

- **الگوهای دسترسی غیرمعمول به داده:**
  - ناهنجاری‌ها در دسترسی به فایل‌ها، پرس‌وجوهای پایگاه‌داده یا فراخوانی‌های API که با فعالیت‌های قانونی مرتبط نیستند.
- **افزایش ترافیک شبکه:**
  - حجم ترافیک خروجی بیشتر از حد طبیعی، به‌ویژه به آدرس‌های IP یا دامنه‌های ناشناخته.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربران ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.

- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.
- **تغییرات غیرمنتظره فایل‌ها:**
  - ناهنجاری‌ها در الگوهای ایجاد، حذف یا تغییر فایل‌ها که با عملیات قانونی مرتبط نیستند.

۲-۴۴-۷ تأثیرات

- **نشت داده‌ها:**
  - سرقت اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به ضررهای مالی یا آسیب‌های اعتباری می‌شود.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به‌دلیل اعتبارهای به‌خطرافتاده یا آسیب‌پذیری‌های مورد سوءاستفاده.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، عفونت‌های باج‌افزاری یا اختلال در خدمات که باعث آسیب‌های مالی می‌شوند.
- **جریمه‌های نظارتی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، که منجر به جریمه‌ها و پیامدهای قانونی می‌شود.
- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا فرم‌ورهای حیاتی، که سیستم‌ها را غیرقابل استفاده می‌کند.

۲-۴۴-۸ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به تشخیص الگوهای غیرمعمول دسترسی یا انتقال داده که نشان‌دهنده خروج داده‌ها هستند، باشند.
- **تحلیل رفتاری:**
  - از تحلیل رفتاری مبتنی بر هوش مصنوعی برای نظارت بر فعالیت کاربران و سیستم‌ها برای

انحرافاتی که ممکن است نشان‌دهنده قصد مخرب باشد، استفاده کنید.

- **رمزنگاری داده:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حال انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده را به حداقل برسانید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش دسترسی غیرمجاز را محدود کرده و تأثیر خروج موفق داده‌ها را کاهش دهید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرمور و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده توسط تکنیک‌های خروج داده‌ها را برطرف کنید.
- **پیشگیری از نشت داده (DLP):**
  - راه‌حل‌های DLP را پیاده‌سازی کنید که از هوش مصنوعی برای شناسایی و مسدود کردن انتقال‌های غیرمجاز داده بر اساس محتوا، زمینه یا رفتار استفاده می‌کنند.
- **کنترل‌های دسترسی:**
  - کنترل‌های دسترسی سخت‌گیرانه، احراز هویت چندعاملی و اصل حداقل دسترسی را اعمال کنید تا دسترسی غیرمجاز را محدود کنید.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد شناسایی نشانه‌های خروج داده‌ها، مانند تلاش‌های فیشینگ یا رفتار غیرمعمول سیستم آموزش دهید.
- **هوشمندسازی تهدیدات پیشرفته:**
  - از پلتفرم‌های هوشمند مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور خروج داده‌ها و تهدیدات تطبیقی مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی‌ای توسعه دهید و پیاده‌سازی کنید تا حملات خروج داده‌ها را به‌سرعت شناسایی، مهار و کاهش دهید.

۲-۴۵ تغییر فیلم‌های نظارتی

تغییر فیلم‌های نظارتی[1] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای دستکاری یا تغییر فیلم‌ها و ضبط‌های سیستم‌های نظارتی است. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند تغییرات بسیار واقع‌گرایانه اما مصنوعی در فیلم‌های نظارتی ایجاد کنند، مانند حذف افراد، اضافه کردن اشیاء یا تغییر کامل رویدادها. این نوع حمله به‌خصوص خطرناک است زیرا اعتماد به شواهد بصری را تضعیف می‌کند، که اغلب برای اهداف امنیتی، اجرای قانون یا تحقیقات پزشکی قانونی مورد استفاده قرار می‌گیرد.

تغییر فیلم‌های نظارتی می‌تواند به‌طور مخرب برای پوشش ردپا پس از یک نفوذ فیزیکی، متهم کردن افراد بی‌گناه یا گمراه کردن تحقیقات استفاده شود. خودکارسازی و واقع‌گرایی ارائه‌شده توسط هوش مصنوعی این تهدید را به‌طور فزاینده‌ای در دسترس و تشخیص آن را دشوارتر می‌کند.

۲-۴۵-۱ ویژگی‌های کلیدی

- **واقع‌گرایی بالا:**
  - هوش مصنوعی تغییراتی ایجاد می‌کند که تقریباً از فیلم‌های واقعی غیرقابل تشخیص هستند و برای ناظران انسانی و سیستم‌های خودکار متقاعدکننده هستند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند تحلیل، تغییر و بازگرداندن فیلم‌های دستکاری‌شده به سیستم‌های نظارتی را خودکار می‌کنند.
- **انعطاف‌پذیری:**
  - تغییرات مبتنی بر هوش مصنوعی به‌طور پویا تکامل می‌یابند و با تغییرات در زوایای دوربین، شرایط نور یا عوامل محیطی سازگار می‌شوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تغییرات هدف‌مند، مانند حذف افراد خاص یا تغییر زمان‌بندی‌ها، ایجاد کنند.
- **پنهان‌کاری:**
  - تغییرات ظریف هستند و به‌طور یکپارچه با فیلم اصلی ترکیب می‌شوند و از تشخیص توسط انسان‌ها و ابزارهای تحلیل مبتنی بر هوش مصنوعی اجتناب می‌کنند.

1. Surveillance Footage Modification

۲-۴۵-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری عمیق:**
  - شبکه‌های مولد تخاصمی: تغییرات مصنوعی اما واقع‌گرایانه در فیلم‌های نظارتی ایجاد می‌کنند، مانند جایگزینی چهره‌ها، حذف اشیاء یا تغییر پس‌زمینه‌ها.
  - شبکه‌های عصبی پیچشی: الگوهای بصری پیچیده در فیلم‌ها را تحلیل و درک می‌کنند تا تغییرات را هدایت کنند.
- **پردازش زبان طبیعی:**
  - متاداده‌های متنی مرتبط با فیلم‌های نظارتی، مانند زمان‌بندی‌ها یا داده‌های مکان، را تحلیل می‌کند تا تغییرات را بهبود بخشد و سازگاری را تضمین کند.
- **بینایی کامپیوتری:**
  - از تکنیک‌های پیشرفته تشخیص تصویر برای شناسایی عناصر کلیدی در فیلم‌ها، مانند افراد، وسایل نقلیه یا اقدامات، برای دستکاری هدف‌مند استفاده می‌کند.
- **یادگیری تقویتی:**
  - استراتژی‌های تغییر را با پاداش‌دهی به فریب موفق و جریمه‌کردن شکست‌ها بهینه می‌کند و امکان بهبود مداوم را فراهم می‌کند.
- **تحلیل زمانی:**
  - تحلیل سری‌های زمانی را اعمال می‌کند تا اطمینان حاصل شود که تغییرات با توالی رویدادها در فیلم هم‌خوانی دارند و انسجام را حفظ می‌کنند.

۲-۴۵-۳ دارایی‌های هدف

- **سیستم‌های امنیتی:**
  - سیستم‌های نظارتی شرکتی، دولتی یا عمومی که در آن‌ها فیلم‌های دستکاری‌شده می‌توانند تحقیقات را گمراه کنند یا نفوذها را پوشش دهند.
- **اجرای قانون:**
  - ادارات پلیس یا سیستم‌های قضایی که به فیلم‌های نظارتی به‌عنوان شواهد متکی هستند، جایی که تغییرات می‌تواند منجر به محکومیت‌های نادرست یا از دست رفتن فرصت‌ها شود.
- **زیرساخت‌های حیاتی:**
  - نیروگاه‌ها، بیمارستان‌ها، فرودگاه‌ها یا مراکز حمل‌ونقل که در آن‌ها فیلم‌های دستکاری‌شده

می‌توانند باعث اختلال در عملیات یا خطرات ایمنی شوند.

- **خرده‌فروشی و هتلداری:**
  - کسب‌وکارهایی که از نظارت برای کنترل سرقت، کلاهبرداری یا رفتار مشتری استفاده می‌کنند، جایی که تغییرات می‌تواند منجر به ضررهای مالی یا آسیب به اعتبار شود.
- **حریم خصوصی شخصی:**
  - افرادی که فیلم‌های نظارتی آن‌ها دستکاری می‌شود تا برای جرایم متهم شوند یا حریم خصوصی آن‌ها نقض شود.

۴-۴۵-۲ آسیب‌پذیری‌ها

- **احراز هویت ضعیف:**
  - عدم وجود مکانیسم‌های احراز هویت قوی برای دسترسی یا تغییر فیلم‌های نظارتی، سیستم‌ها را در معرض خطر قرار می‌دهد.
- **ثبت‌نشده‌های ناکافی:**
  - روش‌های ضعیف ثبت‌نشده، ردیابی تغییرات غیرمجاز یا انتساب آن‌ها به بازیگران خاص را دشوار می‌کند.
- **سیستم‌های قدیمی:**
  - استفاده از نرم‌افزار یا سخت‌افزار نظارتی قدیمی که حاوی آسیب‌پذیری‌های شناخته‌شده قابل بهره‌برداری توسط حملات مبتنی بر هوش مصنوعی هستند.
- **خطای انسانی:**
  - سیستم‌های نادرست پیکربندی‌شده، اعتبارنامه‌های مشترک یا آموزش ناکافی به‌طور ناخواسته آسیب‌پذیری‌هایی را ایجاد می‌کنند که در طول تغییر فیلم قابل بهره‌برداری هستند.
- **عدم وجود بررسی‌های یکپارچگی:**
  - عدم وجود مکانیسم‌های هش رمزنگاری یا تأیید مبتنی بر بلاک‌چین، تغییرات کشف‌نشده را تسهیل می‌کند.

۵-۴۵-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری فیلم‌های نظارتی، پیکربندی‌های سیستم یا اعتبارنامه‌های دسترسی از محیط‌های هدف استفاده می‌کند.

- **تحلیل و آموزش:**
  - مدل‌های یادگیری ماشین را برای تحلیل فیلم‌های جمع‌آوری‌شده و آموزش روی الگوهای رفتار قانونی به‌کار می‌گیرد تا استراتژی‌های تغییر را بهبود بخشد.
- **برنامه‌ریزی تغییر:**
  - از هوش مصنوعی برای طراحی تغییرات دقیق، مانند حذف افراد، تغییر زمان‌بندی‌ها یا جعل رویدادها استفاده می‌کند.
- **اجرا:**
  - تغییرات برنامه‌ریزی‌شده را با تزریق فیلم‌های دستکاری‌شده به سیستم نظارتی اجرا می‌کند و اطمینان می‌یابد که با فریم‌های اطراف سازگار باشد.
- **پایداری:**
  - تغییرات را بر اساس بازخورد به‌طور مداوم بهبود می‌بخشد و اثربخشی بلندمدت و فرار از تشخیص را حفظ می‌کند.
- **پاک‌کردن ردپا:**
  - لاگ‌ها را تغییر می‌دهد، ردپاها را پاک می‌کند یا از تکنیک‌های دیگر برای جلوگیری از تشخیص و حفظ پنهان‌کاری استفاده می‌کند.

۶-۴۵-۲ نشانه‌های احتمال نفوذ

- **ناهنجاری‌ها در فیلم:**
  - ناسازگاری‌های ظریف در نور، سایه‌ها یا حرکت اشیاء که از هنجارهای مورد انتظار منحرف می‌شوند.
- **افزایش استفاده از منابع:**
  - مصرف بیشتر CPU، حافظه یا فضای ذخیره‌سازی به دلیل فرآیندهای تغییر خودکار که در پس‌زمینه اجرا می‌شوند.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار عادی کاربر، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرعادی در سیستم‌های نظارتی.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت

ناموفق یا تلاش‌های دسترسی غیرمجاز.

- **به‌روزرسانی‌های غیرمنتظره سیستم:**
  - انحراف در به‌روزرسانی‌ها یا پیکربندی‌های سیستم که ممکن است نشان‌دهنده وجود ابزارهای تغییر غیرمجاز باشد.

۲-۴۵-۷ تأثیرات

- **اختلال در عملیات:**
  - فیلم‌های دستکاری‌شده می‌توانند عملیات امنیتی را مختل کنند و منجر به از دست رفتن تهدیدها یا هشدارهای نادرست شوند.
- **پیامدهای قانونی:**
  - شواهد دستکاری‌شده در پرونده‌های قضایی می‌تواند منجر به محکومیت‌های نادرست یا تبرئه‌های ناعادلانه شود و سیستم‌های عدالت را تضعیف کند.
- **آسیب به اعتبار:**
  - سازمان‌هایی که از فیلم‌های نظارتی دستکاری‌شده رنج می‌برند ممکن است اعتماد مشتریان، شرکا یا ذینفعان را از دست بدهند.
- **ضررهای مالی:**
  - فعالیت‌های غیرمجاز پوشش‌داده‌شده توسط فیلم‌های دستکاری‌شده می‌تواند منجر به کلاهبرداری مالی، سرقت یا اختلال در خدمات شود.
- **خطرات ایمنی:**
  - در زیرساخت‌های حیاتی، فیلم‌های دستکاری‌شده می‌توانند پاسخ به شرایط اضطراری را به تأخیر بیندازند یا با پنهان‌کردن رویدادهای واقعی جان افراد را به خطر بیندازند.

۲-۴۵-۸ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به شناسایی آثار ظریف یا ناسازگاری‌ها در فیلم‌های نظارتی هستند که نشان‌دهنده دستکاری است.
- **بررسی‌های یکپارچگی رمزنگاری:**
  - مکانیسم‌های هش یا تأیید مبتنی بر بلاک‌چین را پیاده‌سازی کنید تا اصالت و یکپارچگی فیلم‌های ضبط‌شده را تضمین کنند.

- **کنترل‌های دسترسی:**
  - کنترل‌های دسترسی سخت‌گیرانه، احراز هویت چندعاملی و اصل حداقل امتیاز را اعمال کنید تا دسترسی غیرمجاز به سیستم‌های نظارتی محدود شود.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های مکرر از سیستم‌های نظارتی، آرشیوهای فیلم و لاگ‌های دسترسی انجام دهید تا فعالیت‌های مشکوک را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **رعایت اصول رمزنگاری:**
  - فیلم‌های حساس را هم در حالت ذخیره‌سازی و هم در حال انتقال رمزنگاری کنید تا ارزش داده‌های سرقت‌شده یا دستکاری‌شده به حداقل برسد.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد شناسایی علائم دستکاری نظارتی و رعایت بهداشت سایبری خوب آموزش دهید.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور تغییر و تهدیدات انطباق‌پذیر مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت تغییرات فیلم‌های نظارتی را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.
- **دفاع پیشگیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های در حال تکامل تغییر استفاده کنید.
- **سیستم‌های مقاوم در برابر دستکاری:**
  - سیستم‌های نظارتی را با ویژگی‌های مقاوم در برابر دستکاری، مانند فرآیندهای بوت ایمن یا رمزنگاری مبتنی بر سخت‌افزار، طراحی کنید تا از تغییرات غیرمجاز جلوگیری شود.

۲-۴۶ نقض حریم خصوصی

حملات نقض حریم خصوصی[1] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای

1. AI- Driven Privacy Attacks

نقض یا تضعیف سیستماتیک حریم خصوصی افراد، سازمان‌ها یا سیستم‌ها است. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند حجم عظیمی از داده‌ها را تحلیل کنند تا اطلاعات حساسی مانند ترجیحات شخصی، جزئیات مالی، سوابق پزشکی یا اسرار شرکتی را استنباط کنند. این حملات از ضعف‌های مکانیسم‌های حفاظت از داده‌ها، رفتار انسانی و پیکربندی سیستم‌ها بهره‌برداری می‌کنند تا از دفاع‌های سنتی مانند رمزنگاری، کنترل‌های دسترسی یا تکنیک‌های ناشناس‌سازی عبور کنند.

این نوع حمله دقت و انعطاف‌پذیری هوش مصنوعی را با ماهیت فراگیر شیوه‌های جمع‌آوری داده‌های مدرن ترکیب می‌کند و به مهاجمان امکان می‌دهد بینش‌های ارزشمندی را از مجموعه‌داده‌هایی که به ظاهر بی‌ضرر هستند استخراج کنند. حملات نقض حریم خصوصی مبتنی بر هوش مصنوعی می‌توانند منجر به سرقت هویت، کلاهبرداری مالی، آسیب به اعتبار یا حتی خطرات امنیت ملی شوند.

۱-۴۶-۲ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند جمع‌آوری، تحلیل و بهره‌برداری از داده‌ها را خودکار می‌کند و تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی به‌طور پویا تکامل می‌یابند و با تغییرات در اقدامات حریم خصوصی، منابع داده یا شرایط محیطی سازگار می‌شوند.
- **پنهان‌کاری:**
  - با تقلید از ترافیک قانونی یا رمزنگاری payloadها تا زمان اجرا، هوش مصنوعی اطمینان می‌دهد که نقض حریم خصوصی کشف نشود.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند مجموعه‌داده‌ها یا افراد خاص با اطلاعات باارزش را شناسایی کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد هزاران یا میلیون‌ها کاربر را به‌طور همزمان هدف قرار دهند و تأثیر بالقوه را افزایش دهند.

۲-۴۶-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های برچسب‌دار از نقض‌های شناخته‌شده حریم خصوصی آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا خوشه‌ها در داده‌های بدون ساختار را شناسایی می‌کند تا روابط یا آسیب‌پذیری‌های پنهان را آشکار کند.
  - یادگیری تقویتی: تاکتیک‌های حمله را با پاداش‌دهی به نتایج موفق و جریمه‌کردن شکست‌ها بهینه می‌کند و امکان بهبود مداوم را فراهم می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در داده‌ها را تحلیل می‌کنند تا اطلاعات حساس را استنباط کنند.
  - شبکه‌های مولد تخاصمی: مجموعه‌داده‌های مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های فرار را آزمایش و بهبود بخشند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند ایمیل‌ها، پست‌های شبکه‌های اجتماعی یا سوابق عمومی، را تحلیل می‌کند تا بینش‌ها، احساسات یا اطلاعات شخصی قابل شناسایی (PII) را استخراج کند.
- **داده‌کاوی:**
  - از هوش مصنوعی برای استخراج داده‌های عمومی از شبکه‌های اجتماعی، فروم‌ها یا پایگاه‌داده‌های نفوذشده استفاده می‌کند تا پروفایل‌های دقیقی از اهداف بالقوه ایجاد کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا پیش‌بینی کند چه زمانی و چگونه آسیب‌پذیری‌ها را بهره‌برداری کند.

۲-۴۶-۳ دارایی‌های هدف

- **داده‌های شخصی:**
  - اطلاعات حساس درباره افراد، مانند شماره‌های تأمین اجتماعی، سوابق پزشکی یا جزئیات مالی.
- **سیستم‌های سازمانی:**
  - شبکه‌های شرکتی که مالکیت فکری، اسرار تجاری یا داده‌های مشتریان را ذخیره می‌کنند.

- **سوابق بهداشتی:**
  - سوابق الکترونیک سلامت (EHRs)[1] یا لاگ‌های دستگاه‌های پزشکی که نقض حریم خصوصی در آن‌ها می‌تواند ایمنی بیمار یا محرمانه‌بودن را به خطر بیندازد.
- **مؤسسات مالی:**
  - بانک‌ها، اتحادیه‌های اعتباری یا پردازنده‌های پرداخت که در آن‌ها نقض حریم خصوصی می‌تواند منجر به انتقال غیرمجاز وجوه یا کلاهبرداری شود.
- **سازمان‌های دولتی:**
  - اهداف باارزش که اطلاعات طبقه‌بندی‌شده یا زیرساخت‌های حیاتی را کنترل می‌کنند.

۴-۴۶-۲ آسیب‌پذیری‌ها

- **رمزنگاری ضعیف:**
  - استفاده از پروتکل‌های رمزنگاری قدیمی یا نادرست، داده‌های حساس را در معرض دسترسی غیرمجاز قرار می‌دهد.
- **ناشناس‌سازی ناکافی:**
  - مجموعه‌داده‌های ناشناس‌سازی‌شده ضعیف، افراد را در برابر شناسایی مجدد از طریق تحلیل مبتنی بر هوش مصنوعی آسیب‌پذیر می‌کنند.
- **خطای انسانی:**
  - کارکنانی که اطلاعات شخصی را در شبکه‌های اجتماعی بیش از حد به اشتراک می‌گذارند یا در دام کلاهبرداری فیشینگ می‌افتند، به‌طور ناخواسته حریم خصوصی خود را در معرض خطر قرار می‌دهند.
- **دفاع‌های قدیمی:**
  - سیستم‌های قدیمی یا نرم‌افزارهای منسوخ، سازمان‌ها را در برابر حملات پیشرفته نقض حریم خصوصی آسیب‌پذیر می‌کنند.
- **اتکای بیش از حد به خودکارسازی:**
  - سیستم‌های تشخیص سنتی ممکن است نتوانند تهدیدات ظریف یا در حال تکامل تولیدشده توسط حملات نقض حریم خصوصی مبتنی بر هوش مصنوعی را تشخیص دهند.

1. Electronic Health Record (EHR)

۵-۴۶-۲ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری داده‌های عمومی، ارتباطات رهگیری‌شده یا مجموعه‌داده‌های نشت‌یافته استفاده می‌کند.
- **تحلیل و آموزش:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های جمع‌آوری‌شده و آموزش روی الگوهای رفتار قانونی به‌کار می‌گیرد تا استراتژی‌های بهره‌برداری را بهبود بخشد.
- **برنامه‌ریزی نقض حریم خصوصی:**
  - مهاجم از هوش مصنوعی برای شناسایی اهداف باارزش، مانند مدیران، کارکنان یا سیستم‌های ذخیره‌سازی داده‌های حساس استفاده می‌کند.
- **اجرا:**
  - مهاجم از آسیب‌پذیری‌های شناسایی‌شده برای استخراج اطلاعات حساس بهره‌برداری می‌کند، اغلب بدون ایجاد هشدار یا باقی‌گذاشتن ردپا.
- **پایداری:**
  - مهاجم رویکرد حمله را بر اساس بازخورد به‌طور مداوم بهبود می‌بخشد تا دسترسی بلندمدت یا تأثیر بر سیستم‌های هدف را حفظ کند.
- **سوءاستفاده:**
  - مهاجم می‌تواند داده‌های سرقت‌شده را در دارک وب بفروشد، از آن برای باج‌گیری استفاده کند یا از آن برای حملات بیشتر مانند فیشینگ هدفمند یا سرقت هویت بهره‌برداری کند.

۶-۴۶-۲ نشانه‌های احتمال نفوذ

- **الگوهای غیرعادی دسترسی به داده:**
  - ناهنجاری‌ها در دسترسی به فایل‌ها، پرس‌وجوهای پایگاه داده یا فراخوانی‌های API که با فعالیت‌های قانونی مرتبط نیستند.
- **افزایش استفاده از منابع:**
  - فعالیت بیشتر CPU، حافظه یا I/O دیسک به دلیل تحلیل یا استخراج خودکار داده.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار عادی کاربر، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای

غیرعادی.

- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.
- **ترافیک شبکه غیرمنتظره:**
  - انحراف در ترافیک خروجی، مانند اتصال به آدرس‌های IP یا دامنه‌های ناشناخته، که ممکن است نشان‌دهنده خروج داده باشد.

۲-۴۶-۷ تأثیرات

- **سرقت هویت:**
  - دسترسی غیرمجاز به اطلاعات شخصی که منجر به سوءاستفاده از اعتبارنامه یا جعل هویت می‌شود.
- **ضررهای مالی:**
  - تراکنش‌های کلاهبرداری، تصاحب حساب یا سایر اشکال بهره‌برداری مالی که باعث آسیب مالی می‌شوند.
- **آسیب به اعتبار:**
  - سازمان‌هایی که از نقض حریم خصوصی رنج می‌برند ممکن است اعتماد مشتریان و ارزش برند خود را از دست بدهند.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل اعتبارنامه‌های به‌خطرافتاده یا آسیب‌پذیری‌های سوءاستفاده‌شده.
- **جریمه‌های نظارتی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، که منجر به جریمه‌ها و اقدامات قانونی می‌شود.
- **خطرات ایمنی:**
  - در بخش بهداشت یا زیرساخت‌های حیاتی، نقض حریم خصوصی می‌تواند جان افراد را به خطر بیندازد یا باعث شکست‌های فاجعه‌بار شود.

۲-۴۶-۸ راه‌های مقابله

- **کاهش داده:**
  - فقط حداقل داده‌های لازم را جمع‌آوری و ذخیره کنید تا تأثیر احتمالی نقض به حداقل برسد.

- **رعایت اصول رمزنگاری:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حال انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده به حداقل برسد.
- **تکنیک‌های ناشناس‌سازی:**
  - روش‌های قوی ناشناس‌سازی، مانند حریم خصوصی تفاضلی یا k-ناشناس‌سازی، را پیاده‌سازی کنید تا هویت افراد در مجموعه‌داده‌ها محافظت شود.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به تشخیص الگوهای غیرعادی دسترسی یا انتقال داده هستند که نشان‌دهنده نقض حریم خصوصی است.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای نظارت بر فعالیت کاربر و سیستم برای انحرافاتی که ممکن است نشان‌دهنده قصد مخرب باشد، استفاده کنید.
- **کنترل‌های دسترسی:**
  - کنترل‌های دسترسی سخت‌گیرانه، احراز هویت چندعاملی و اصل حداقل امتیاز را اعمال کنید تا دسترسی غیرمجاز محدود شود.
- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد شناسایی علائم نقض حریم خصوصی آموزش دهید و فرهنگ هوشیاری را تقویت کنید.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور حملات نقض حریم خصوصی و تهدیدات انطباق‌پذیر مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حملات نقض حریم خصوصی را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.

۲-۴۷ هماهنگی حملات

هماهنگی حملات[1] مبتنی بر هوش مصنوعی شامل استفاده از هوش مصنوعی برای هماهنگی و همگام‌سازی چندین حمله سایبری در سیستم‌ها، شبکه‌ها یا پلتفرم‌های مختلف است. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند برنامه‌ریزی، اجرا و تطبیق حملات چندبردار پیچیده را خودکار کنند و در عین حال تأثیر را به حداکثر برسانند و خطر تشخیص را به حداقل برسانند. این نوع حمله به‌ویژه خطرناک است زیرا به مهاجمان امکان می‌دهد تا تکنیک‌های مختلفی مانند فیشینگ، استقرار بدافزار، DDoS یا حرکت جانبی را در یک کمپین منسجم ترکیب کنند که دفاع‌های سنتی را تحت فشار قرار می‌دهد.

هماهنگی حملات مبتنی بر هوش مصنوعی دقت، انعطاف‌پذیری و مقیاس‌پذیری عملیات سایبری را افزایش می‌دهد و پاسخ‌گویی مؤثر مدافعان را دشوارتر می‌کند. همچنین به مهاجمان امکان می‌دهد تا استراتژی‌های خود را بر اساس بازخوردهای لحظه‌ای به‌طور پویا تنظیم کنند و موفقیت پایدار را تضمین کنند.

۲-۴۷-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند شناسایی اهداف، انتخاب بردارهای حمله و هماهنگی اقدامات در چندین سیستم را خودکار می‌کند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی به‌طور لحظه‌ای تکامل می‌یابند و با تغییرات در اقدامات دفاعی، پیکربندی سیستم‌ها یا شرایط محیطی سازگار می‌شوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا آسیب‌پذیری‌ها، سیستم‌ها یا افراد خاص را برای حداکثر تأثیر شناسایی کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا کمپین‌های بزرگ‌مقیاسی را هماهنگ کنند که هزاران یا میلیون‌ها سیستم را به‌طور همزمان هدف قرار می‌دهند.
- **پنهان‌کاری:**
  - با تقلید از فعالیت‌های قانونی یا رمزنگاری payloadها تا زمان اجرا، هوش مصنوعی اطمینان

1. Attack Cordination

می‌دهد که حملات هماهنگ‌شده بدون تشخیص باقی بمانند.

۲-۴۷-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های موفق و ناموفق تلاش‌های حمله آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا ناهنجاری‌ها را در ترافیک شبکه، رفتار کاربر یا لاگ‌های سیستم شناسایی می‌کند تا نقاط ضعف بالقوه را آشکار کند.
  - یادگیری تقویتی: هماهنگی حملات را با پاداش دادن به نتایج موفق و جریمه کردن شکست‌ها بهینه می‌کند، که امکان بهبود مستمر را فراهم می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، وابستگی‌های زمانی در تعاملات سیستم را تحلیل می‌کنند تا زمان‌بندی و توالی حملات را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: سناریوهای حمله مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های فرار را آزمایش و بهبود بخشند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار طبیعی کاربر یا سیستم استفاده می‌کند تا مهاجمان بتوانند فعالیت‌های مخرب را با فعالیت‌های قانونی ترکیب کنند.
- **هوش جمعی:**
  - از تکنیک‌های هوش مصنوعی غیرمتمرکز برای هماهنگی اقدامات بین چندین دستگاه یا سیستم نفوذشده استفاده می‌کند و رفتار جمعی را تقلید می‌کند.
- **داده‌کاوی:**
  - حجم عظیمی از داده‌ها را از سیستم‌های نفوذشده تحلیل می‌کند تا اعتبارنامه‌ها، آسیب‌پذیری‌ها یا روابط مورد اعتماد را کشف کند که هماهنگی را تسهیل می‌کنند.

۲-۴۷-۳ دارایی‌های هدف

- **شبکه‌های سازمانی:**
  - اینترانت‌های شرکتی که حملات هماهنگ‌شده می‌توانند منجر به دسترسی غیرمجاز به داده‌های حساس یا سیستم‌های حیاتی شوند.

- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، مراکز بهداشتی، مؤسسات مالی یا شبکه‌های حمل‌ونقل که حملات هماهنگ‌شده می‌توانند آسیب‌های جدی ایجاد کنند.
- **خدمات ابری:**
  - برنامه‌ها، ذخیره‌سازی یا منابع محاسباتی میزبانی‌شده در ابر که از طریق اعتبارنامه‌های نفوذشده یا پیکربندی‌های نادرست قابل دسترسی هستند.
- **شبکه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های اینترنت اشیا با منابع محدود که اغلب به عنوان بخشی از بات‌نت‌ها یا نقاط ورودی برای حملات هماهنگ‌شده استفاده می‌شوند.
- **سیستم‌های زنجیره تأمین:**
  - تأمین‌کنندگان، شرکا یا فروشندگان شخص ثالث که حساب‌های آن‌ها می‌تواند به عنوان پله‌هایی برای کمپین‌های بزرگتر مورد سوءاستفاده قرار گیرد.

۴-۴۷-۲ آسیب‌پذیری‌ها

- **تقسیم‌بندی ضعیف شبکه:**
  - نبود تقسیم‌بندی قوی، سیستم‌ها را در برابر عبور و سوءاستفاده هماهنگ آسیب‌پذیر می‌کند.
- **نظارت ناکافی:**
  - روش‌های ضعیف ثبت لاگ یا نبود نظارت لحظه‌ای، تشخیص فعالیت‌های هماهنگ‌شده را دشوار می‌کند.
- **خطای انسانی:**
  - فایروال‌های نادرست پیکربندی‌شده، اعتبارنامه‌های مشترک یا آموزش ناکافی به‌طور ناخواسته آسیب‌پذیری‌هایی را در معرض سوءاستفاده در طول هماهنگی قرار می‌دهند.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های قدیمی یا نرم‌افزارهای منسوخ، سازمان‌ها را در برابر تکنیک‌های هماهنگی پیشرفته آسیب‌پذیر می‌کند.
- **عدم وجود اقدامات متقابل:**
  - نبود دفاع‌های مبتنی بر هوش مصنوعی یا هوش تهدید پیش‌گیرانه، سیستم‌ها را در برابر

استراتژی‌های هماهنگی تطبیقی آسیب‌پذیر می‌کند.

۲-۴۷-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره سیستم‌های هدف، از جمله معماری، پیکربندی‌ها و رفتارهای معمول آن‌ها استفاده می‌کند.
- **انتخاب هدف:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های جمع‌آوری‌شده و شناسایی اهداف باارزش یا نقاط ضعف برای حمله مستقر می‌کند.
- **برنامه‌ریزی حمله:**
  - مهاجم از هوش مصنوعی برای طراحی و ترتیب حملات استفاده کنید و تخصیص منابع و زمان‌بندی را برای حداکثر تأثیر بهینه‌سازی می‌کند.
- **استقرار:**
  - مهاجم حملات برنامه‌ریزی‌شده را در چندین سیستم یا شبکه اجرا کرده و اطمینان حاصل می‌کند که با زمینه قربانی هماهنگ هستند و خطر تشخیص را به حداقل می‌رسانند.
- **هماهنگی:**
  - مهاجم اقدامات بین دستگاه‌ها، سیستم‌ها یا حساب‌های نفوذشده را همگام‌سازی می‌کند تا اهدافی مانند استخراج داده، استقرار باج‌افزار یا اختلال خدمات محقق شوند.
- **پایداری:**
  - مهاجم کمپین را به‌طور مداوم بر اساس بازخورد بهبود می‌بخشد تا دسترسی بلندمدت یا تأثیر بر محیط‌های هدف حفظ شود.

۲-۴۷-۶ نشانه‌های احتمال نفوذ

- **الگوهای ترافیکی غیرمعمول:**
  - ناهنجاری‌ها در ترافیک شبکه، مانند اتصالات همزمان از چندین سیستم یا جریان‌های داده غیرمنتظره.
- **افزایش استفاده از منابع:**
  - مصرف بیشتر CPU، حافظه یا پهنای باند به دلیل فعالیت‌های هماهنگی خودکار.

- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربر یا سیستم ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.
- **تعاملات غیرمنتظره سیستم:**
  - انحراف در تعاملات سیستم، مانند ارتباط بین سیستم‌هایی که معمولاً با هم تعامل ندارند.

۲-۴۷-۷ تأثیرات

- **اختلال عملیاتی:**
  - حملات هماهنگ‌شده می‌توانند خدمات یا سیستم‌های حیاتی کسب‌وکار را تحت فشار قرار دهند و باعث downtime یا کاهش عملکرد شوند.
- **نقض داده‌ها:**
  - دسترسی غیرمجاز به اطلاعات حساس، از جمله داده‌های شخصی، مالی یا مالکیت فکری، که منجر به ضررهای مالی یا آسیب شهرت می‌شود.
- **ضررهای مالی:**
  - استقرار باج‌افزار، کلاهبرداری یا اختلال خدمات باعث آسیب‌های مالی می‌شود.
- **آسیب به شهرت:**
  - قربانیان حملات هماهنگ‌شده ممکن است اعتماد مشتریان و ارزش برند خود را از دست بدهند.
- **فساد سیستم:**
  - تخریب یا تغییر فایل‌ها، پیکربندی‌ها یا فرم‌ورهای حیاتی، که سیستم‌ها را غیرقابل استفاده می‌کند.

۲-۴۷-۸ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به تشخیص الگوهای ترافیکی غیرعادی یا فعالیت‌های هماهنگ‌شده نشان‌دهنده حملات چندبردار هستند.

- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش دسترسی غیرمجاز یا فعالیت‌های مخرب محدود شود.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای نظارت بر انحرافات در فعالیت سیستم یا کاربر که ممکن است نشان‌دهنده حملات هماهنگ‌شده باشد استفاده کنید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرم‌ورها و پیکربندی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده در طول هماهنگی اصلاح شوند.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا درباره تکنیک‌های نوظهور هماهنگی و تهدیدات تطبیقی به‌روز بمانید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حملات هماهنگ‌شده را تشخیص، مهار و کاهش دهید.
- **دفاع چندلایه:**
  - دفاع‌های سنتی مانند فایروال‌ها، سیستم‌های تشخیص نفوذ (IDS) و محافظت از نقاط پایانی را با راه‌حل‌های مبتنی بر هوش مصنوعی ترکیب کنید تا پوشش جامع فراهم شود.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد تشخیص علائم نفوذ، مانند رفتار غیرمعمول سیستم یا درخواست‌های غیرمنتظره آموزش دهید.
- **دفاع پیش‌گیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های در حال تحول هماهنگی استفاده کنید.
- **دفاع مشارکتی:**
  - همکاری بین سازمان‌ها، دولت‌ها و متخصصان امنیت سایبری را تقویت کنید تا اطلاعات را به اشتراک بگذارند و به‌طور جمعی با تهدیدات هماهنگ مقابله کنند.

۲-۴۸ سیاه‌چاله

حمله سیاه‌چاله[1] نوعی حمله شبکه‌ای است که در آن یک مهاجم بسته‌های ارسالی به گره‌ها یا سیستم‌های خاص را رهگیری و دور می‌اندازد و به‌طور مؤثری یک «سیاه‌چاله» در شبکه ایجاد می‌کند. این کار ارتباط بین موجودیت‌های قانونی را مختل می‌کند و منجر به قطع سرویس، از دست‌دادن داده‌ها یا تغییر مسیر ترافیک از طریق مسیرهای مخرب می‌شود. در حمله سیاه‌چاله مبتنی بر هوش مصنوعی، از هوش مصنوعی برای افزایش دقت، انعطاف‌پذیری و پنهان‌کاری حملات سیاه‌چاله سنتی استفاده می‌شود. با استفاده از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند به‌طور پویا اهداف باارزش را شناسایی کنند، استراتژی‌های رهگیری بسته‌ها را بهینه کنند و از مکانیزم‌های تشخیص فرار کنند.

حملات سیاه‌چاله مبتنی بر هوش مصنوعی به‌ویژه در شبکه‌های غیرمتمرکز مانند شبکه‌های ad hoc موبایل (MANETs)، شبکه‌های حسگر بی‌سیم (WSNs) و سیستم‌های همتا به همتا (P2P) مبتنی بر بلاک‌چین خطرناک هستند، جایی که اعتماد و قابلیت اطمینان برای ارتباطات حیاتی هستند.

۲-۴۸-۱ ویژگی‌های کلیدی

- **انتخاب پویای اهداف:**
  - هوش مصنوعی گره‌های باارزش یا مسیرهای ارتباطی حیاتی را برای ایجاد حداکثر اختلال شناسایی می‌کند.
- **خودکارسازی:**
  - الگوریتم‌های یادگیری ماشین فرآیند رهگیری و دورانداختن بسته‌ها را خودکار می‌کنند و تلاش دستی را کاهش می‌دهند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با تغییرات در توپولوژی شبکه، پروتکل‌های مسیریابی یا مکانیزم‌های دفاعی سازگار شوند.
- **پنهان‌کاری:**
  - با تحلیل الگوهای ترافیک عادی، هوش مصنوعی اطمینان می‌دهد که حمله به‌راحتی با فعالیت‌های قانونی ترکیب شود و از تشخیص توسط سیستم‌های تشخیص نفوذ (IDS) جلوگیری کند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین گره یا مسیر را به‌طور همزمان هدف

1. Black Hole Attack

قرار دهند و دامنه و تأثیر حمله را افزایش دهند.

۲-۴۸-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های ترافیک شبکه آموزش می‌دهد تا پیش‌بینی کند کدام بسته‌ها یا گره‌ها بیشترین ارزش را برای حمله دارند.
  - یادگیری بدون نظارت: ناهنجاری‌ها یا الگوهای رفتار شبکه را شناسایی می‌کند تا انتخاب اهداف را بدون برچسب‌گذاری قبلی هدایت کند.
  - یادگیری تقویتی: استراتژی‌های حمله را با پاداش دادن به رهگیری موفق بسته‌ها و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و شبکه‌های حافظه کوتاه‌مدت بلندمدت، وابستگی‌های زمانی پیچیده در ترافیک شبکه را تحلیل می‌کنند تا الگوهای ارتباطی آینده را پیش‌بینی کنند.
  - شبکه‌های مولد تخاصمی: الگوهای ترافیکی مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا استحکام اقدامات دفاعی را آزمایش کنند یا حمله را پنهان کنند.
- **شبکه‌های عصبی گراف (GNNs):**
  - برای مدل‌سازی و تحلیل ساختار توپولوژی‌های شبکه استفاده می‌شوند و به مهاجمان امکان می‌دهند گره‌ها یا مسیرهای حیاتی را برای ایجاد اختلال شناسایی کنند.
- **پردازش زبان طبیعی:**
  - متادیتای متنی، مانند فایل‌های لاگ یا جداول مسیریابی، را تحلیل می‌کند تا زمینه‌های بیشتری درباره شبکه استنباط کند.
- **تحلیل سری‌های زمانی:**
  - از روش‌های آماری و یادگیری ماشین برای تحلیل وابستگی‌های زمانی در ترافیک شبکه استفاده می‌کند و به پیش‌بینی رفتارهای آینده کمک می‌کند.

۲-۴۸-۳ دارایی‌های هدف

- **شبکه‌های ad hoc موبایل (MANETs):**
  - شبکه‌های غیرمتمرکز که در آن گره‌ها برای مسیریابی به یکدیگر متکی هستند و در برابر

حملات سیاه‌چاله آسیب‌پذیر هستند.

- **شبکه‌های حسگر بی‌سیم (WSNs):**
  - دستگاه‌های با منابع محدود که به‌طور متناوب ارتباط برقرار می‌کنند و فرصت‌هایی برای مهاجمان برای رهگیری و دورانداختن بسته‌ها فراهم می‌کنند.
- **شبکه‌های بلاک‌چین:**
  - شبکه‌های همتا به همتا مورد استفاده در سیستم‌های بلاک‌چین، جایی که اختلال در ارتباطات می‌تواند منجر به شکست اجماع یا تأخیر در تراکنش‌ها شود.
- **شبکه‌های سازمانی:**
  - شبکه‌های شرکتی با مسیرهای ارتباطی حیاتی که در صورت اختلال می‌توانند باعث downtime عملیاتی یا ضررهای مالی شوند.
- **شبکه‌های اینترنت اشیا (IoT):**
  - اکوسیستم‌های توزیع‌شده اینترنت اشیا که در آن دستگاه‌های به‌خطر افتاده می‌توانند به‌عنوان سیاه‌چاله عمل کنند و اختلالات گسترده ایجاد کنند.
- **شبکه‌های ارتباطی:**
  - شبکه‌های موبایل، ارائه‌دهندگان خدمات اینترنت (ISPs) و ارائه‌دهندگان VoIP که زیرساخت آن‌ها حجم عظیمی از داده‌های کاربری را مدیریت می‌کنند.

۴-۴۸-۲ آسیب‌پذیری‌ها

- **پروتکل‌های مسیریابی ضعیف:**
  - الگوریتم‌های مسیریابی که صحت اطلاعات مسیریابی را تأیید نمی‌کنند، در برابر حملات سیاه‌چاله آسیب‌پذیر هستند.
- **عدم وجود احراز هویت:**
  - عدم وجود مکانیزم‌های احراز هویت قوی برای به‌روزرسانی‌های مسیریابی، امکان تزریق اطلاعات نادرست را به مهاجمان می‌دهد.
- **نظارت ناکافی:**
  - روش‌های ضعیف نظارت، تشخیص و پاسخ به حملات سیاه‌چاله را در زمان واقعی دشوار می‌کند.

- **خطای انسانی:**
  - تنظیمات نادرست فایروال‌ها، روترها یا سوئیچ‌ها ممکن است به‌طور ناخواسته فرصت‌هایی برای مهاجمان ایجاد کنند.
- **محدودیت منابع:**
  - دستگاه‌هایی با قدرت محاسباتی یا حافظه محدود ممکن است در پیاده‌سازی اقدامات امنیتی قوی مشکل داشته باشند.

۵-۴۸-۲ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای نقشه‌برداری از توپولوژی شبکه استفاده کرده و گره‌های حیاتی، مسیرهای ارتباطی و اهداف بالقوه را شناسایی می‌کند.
- **جمع‌آوری داده:**
  - مهاجم ابزارهای جاسوسی ترافیک شبکه (sniffing) تقویت‌شده توسط یادگیری ماشین را برای ضبط متادیتا از ترافیک شبکه، از جمله اندازه بسته‌ها، زمان‌بندی‌ها و جهت‌های جریان مستقر می‌کند.
- **تحلیل:**
  - مهاجم از الگوریتم‌های یادگیری ماشین برای تحلیل داده‌های جمع‌آوری‌شده استفاده می‌کند و الگوها را شناسایی کرده و فعالیت‌های کاربران یا وضعیت سیستم را استنباط می‌کند.
- **تنظیم رهگیری:**
  - مهاجم، گره موردنظر و مهاجم را به‌طور استراتژیک در شبکه قرار داده تا بسته‌های ارسالی به اهداف خاص را رهگیری کند.
- **دورانداختن بسته‌ها::**
  - مهاجم بسته‌های رهگیری‌شده را بر اساس معیارهای از پیش تعریف‌شده، مانند آدرس‌های IP مقصد یا انواع محتوا، به‌طور پویا دور می‌اندازد.
- **پایداری:**
  - مهاجم به‌طور مداوم شبکه را نظارت می‌کند و استراتژی حمله را بر اساس نتایج مشاهده‌شده اصلاح کرده تا اختلال بلندمدت ایجاد شود.

۲-۴۸-۶ نشانه‌های احتمال نفوذ

- **الگوهای ترافیکی غیرعادی:**
  - ناهنجاری‌ها در نرخ تحویل بسته‌ها، افزایش تأخیر یا مسیرهای مسیریابی غیرمنتظره.
- **افزایش از دست‌دادن بسته‌ها:**
  - افزایش قابل توجه در از دست‌دادن بسته‌ها برای گره‌ها یا مسیرهای ارتباطی خاص.
- **تغییرات ناگهانی در توپولوژی شبکه:**
  - تغییرات غیرمنتظره در جداول مسیریابی یا اتصال گره‌ها که با فعالیت مشکوک همبستگی دارند.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول شبکه، مانند اتصال‌های ناموفق مکرر یا استفاده غیرعادی از منابع.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مرتبط با به‌روزرسانی‌های مسیریابی، بسته‌های دورانداخته‌شده یا فعالیت غیرعادی.

۲-۴۸-۷ تأثیرات

- **اختلال در سرویس:**
  - اختلال در ارتباط بین گره‌های قانونی، منجر به قطع سرویس یا کاهش عملکرد می‌شود.
- **از دست‌دادن داده‌ها:**
  - بسته‌های حیاتی حاوی اطلاعات حساس ممکن است دور انداخته شوند و منجر به از دست‌دادن یا خرابی داده‌ها شوند.
- **Downtime عملیاتی:**
  - اختلال در برنامه‌ها یا سرویس‌های حیاتی کسب‌وکار، باعث ضررهای مالی یا آسیب به اعتبار می‌شود.
- **خطرات امنیتی:**
  - افشای آسیب‌پذیری‌ها در پروتکل‌های مسیریابی یا پیکربندی شبکه، امکان بهره‌برداری بیشتر را فراهم می‌کند.

- **جریمه‌های قانونی:**
  - عدم رعایت توافق‌نامه‌های سطح سرویس (SLAs) یا الزامات قانونی به دلیل قطعی‌های طولانی‌مدت.

۸-۴۸-۲ راه‌های مقابله

- **پروتکل‌های مسیریابی امن:**
  - پروتکل‌های مسیریابی امن، مانند AODV-Sec یا GRP، را پیاده‌سازی کنید که به‌روزرسانی‌های مسیریابی را احراز هویت می‌کنند و از تغییرات غیرمجاز جلوگیری می‌کنند.
- **مکانیزم‌های احراز هویت:**
  - از امضاهای دیجیتال، گواهی‌ها یا سایر تکنیک‌های رمزنگاری برای تأیید صحت اطلاعات مسیریابی استفاده کنید.
- **افزونگی و Failover:**
  - شبکه‌ها را با مسیرهای افزونه و مکانیزم‌های failover طراحی کنید تا حتی در صورت به‌خطر افتادن برخی گره‌ها، ارتباطات ادامه یابد.
- **نظارت رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای ایجاد خطوط پایه و تشخیص انحرافات نشان‌دهنده حملات سیاه‌چاله استفاده کنید.
- **بررسی‌های منظم:**
  - ممیزی‌های امنیتی مکرر انجام دهید تا آسیب‌پذیری‌ها را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **رمزنگاری:**
  - تمام ارتباطات، از جمله متادیتا، را با استفاده از پروتکل‌های رمزنگاری قوی مانند TLS 1.3 یا IPSec رمزنگاری کنید.
- **کنترل‌های دسترسی:**
  - دسترسی به جداول مسیریابی و سیستم‌های مدیریت شبکه را به پرسنل مجاز محدود کنید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا تأثیر حمله سیاه‌چاله محدود شود و جداسازی گره‌های به‌خطر افتاده آسان‌تر شود.

۲-۴۹ دستکاری رکوردها

دستکاری رکوردها[1] شامل تغییر، حذف یا جعل داده‌های ذخیره‌شده در سیستم‌ها، پایگاه‌داده‌ها یا لاگ‌ها به‌منظور گمراه‌کردن، اختلال یا پنهان‌کردن فعالیت‌های مخرب است. در یک حمله دستکاری رکوردها مبتنی بر هوش مصنوعی، از هوش مصنوعی برای خودکارسازی و بهبود فرآیند شناسایی، دسترسی و تغییر رکوردها در حالی که از تشخیص جلوگیری می‌شود، استفاده می‌شود. با استفاده از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند حجم عظیمی از داده‌ها را تحلیل کنند، رفتار سیستم‌ها را پیش‌بینی کنند و تغییراتی ایجاد کنند که به‌طور یکپارچه با ورودی‌های قانونی ترکیب شوند.

این نوع حمله یکپارچگی سیستم‌های حیاتی را تضعیف می‌کند و منجر به بی‌اعتمادی، اختلال در عملیات یا ضررهای مالی می‌شود. دستکاری رکوردها مبتنی بر هوش مصنوعی می‌تواند دفاع‌های سنتی مانند کنترل‌های دسترسی، لاگ‌های حسابرسی یا سیستم‌های تشخیص ناهنجاری را با سازگاری با مکانیزم‌های آن‌ها و تقلید از رفتار عادی دور بزند.

۲-۴۹-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند شناسایی آسیب‌پذیری‌ها، دسترسی به رکوردها و اجرای دستکاری را خودکار می‌کند، که تلاش دستی را کاهش داده و کارایی را افزایش می‌دهد.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا رکوردها یا مجموعه‌داده‌های خاص را با دقت بالا برای تغییر شناسایی کنند.
- **انطباق‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در اقدامات امنیتی، پیکربندی سیستم یا شرایط محیطی سازگار شوند.
- **پنهان‌کاری:**
  - با تقلید از فعالیت‌های قانونی یا رمزنگاری payloadها تا زمان اجرا، هوش مصنوعی اطمینان می‌دهد که دستکاری شناسایی نشود.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد تا چندین سیستم یا مجموعه‌داده‌ها را به‌طور

1. Record Tampering

همزمان هدف قرار دهند، که پتانسیل آسیب را افزایش می‌دهد.

۲-۴۹-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های برچسب‌دار از تکنیک‌های شناخته‌شده دستکاری و نتایج موفق آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در الگوهای داده را شناسایی می‌کند تا فرصت‌های دستکاری را آشکار کند.
  - یادگیری تقویتی: استراتژی‌های دستکاری را با پاداش دادن به تغییرات موفق و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و ترنسفورمرها، وابستگی‌های زمانی در الگوهای دسترسی و تغییر داده‌ها را تحلیل می‌کنند تا واقع‌گرایی را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: ورودی‌های داده مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های دور زدن را تست و بهبود بخشند.
- **پردازش زبان طبیعی:**
  - داده‌های متنی در لاگ‌ها، اسناد یا ارتباطات را تحلیل می‌کند تا زمینه‌های بیشتری درباره اهداف بالقوه یا اقدامات دفاعی استنباط کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند تا تغییرات قانونی را به‌طور متقاعدکننده تقلید کند.
- **داده‌کاوی:**
  - از هوش مصنوعی برای استخراج داده‌های عمومی یا سوءاستفاده از مجموعه‌داده‌های داخلی استفاده می‌کند تا نقاط ضعف در سیستم‌های مدیریت رکوردها را شناسایی کند.

۲-۴۹-۳ دارایی‌های هدف

- **سیستم‌های مالی:**
  - دفترهای بانکی، لاگ‌های تراکنش یا حساب‌های مشتری که در آن‌ها دستکاری می‌تواند منجر به انتقال غیرمجاز وجوه یا کلاهبرداری شود.

- **سوابق بهداشتی:**
  - سوابق الکترونیک سلامت (EHRs) یا لاگ‌های دستگاه‌های پزشکی که در آن‌ها دستکاری می‌تواند ایمنی بیمار یا برنامه‌های درمانی را به خطر بیندازد.
- **اسناد حقوقی:**
  - قراردادها، توافق‌نامه‌ها یا سوابق رسمی که به‌صورت دیجیتالی ذخیره شده‌اند و تغییرات می‌توانند پیامدهای حقوقی یا مالی داشته باشند.
- **پایگاه‌داده‌های سازمانی:**
  - پایگاه‌داده‌های شرکتی که اطلاعات حساس، از جمله سوابق کارمندان، برنامه‌های استراتژیک یا جزئیات عملیاتی را ذخیره می‌کنند.
- **زیرساخت‌های حیاتی:**
  - سیستم‌هایی که شبکه‌های برق، شبکه‌های حمل‌ونقل یا فرآیندهای صنعتی را کنترل می‌کنند و در آن‌ها دستکاری می‌تواند آسیب‌های قابل توجهی ایجاد کند.

۴-۴۹-۲ آسیب‌پذیری‌ها

- **کنترل‌های دسترسی ضعیف:**
  - فقدان مکانیزم‌های احراز هویت یا مجوز قوی، رکوردها را در معرض تغییرات غیرمجاز قرار می‌دهد.
- **لاگ‌گیری ناکافی:**
  - روش‌های ضعیف لاگ‌گیری، ردیابی و نسبت‌دادن فعالیت‌های دستکاری را دشوار می‌کند.
- **نرم‌افزارهای قدیمی:**
  - استفاده از کتابخانه‌ها، فریم‌ورک‌ها یا سیستم‌های عامل قدیمی که حاوی آسیب‌پذیری‌های شناخته‌شده قابل سوءاستفاده توسط تلاش‌های دستکاری هستند.
- **خطای انسانی:**
  - سیستم‌های نادرست پیکربندی‌شده، اعتبارهای مشترک یا آموزش ناکافی به‌طور ناخواسته آسیب‌پذیری‌هایی را ایجاد می‌کنند که توسط مهاجمان مورد سوءاستفاده قرار می‌گیرند.
- **فقدان بررسی‌های یکپارچگی:**
  - فقدان مکانیزم‌هایی مانند هش‌کردن، امضای دیجیتال یا تأیید مبتنی بر بلاک‌چین، دستکاری رکوردها را بدون تشخیص تسهیل می‌کند.

۲-۴۹-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره سیستم هدف، از جمله معماری آن، مکان‌های ذخیره‌سازی داده و الگوهای دسترسی معمول استفاده می‌کند.
- **اسکن آسیب‌پذیری:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل سیستم از نظر ضعف‌هایی مانند کنترل‌های دسترسی نادرست یا رمزنگاری ضعیف مستقر می‌کند.
- **توسعه استراتژی دستکاری:**
  - مهاجم از هوش مصنوعی برای طراحی روش‌های مخفیانه دستکاری که با رفتار عادی سیستم هماهنگ هستند یا از مکانیزم‌های تشخیص فرار می‌کنند، استفاده می‌کند.
- **اجرا:**
  - مهاجم رکوردها را به‌گونه‌ای تغییر می‌دهد که باعث ایجاد هشدار نشود، مانند تغییرات جزئی در مقادیر عددی یا برچسب‌های زمانی.
- **پاک‌کردن ردپاها:**
  - مهاجم لاگ‌ها را تغییر داده، ردپاها را پاک می‌کند یا از تکنیک‌های دیگر برای جلوگیری از تشخیص و حفظ پنهان‌کاری استفاده می‌کند.
- **پایداری:**
  - مهاجم تکنیک‌های دستکاری را بر اساس نتایج مشاهده‌شده به‌طور مداوم بهبود می‌بخشد تا دسترسی یا تأثیر بلندمدت حفظ شود.

۲-۴۹-۶ نشانه‌های احتمال نفوذ

- **تغییرات غیرعادی داده:**
  - تغییرات غیرقابل توضیح در رکوردها، مانند به‌روزرسانی‌ها یا حذف‌های غیرمنتظره، که با عملیات قانونی مرتبط نیستند.
- **افزایش استفاده از منابع:**
  - فعالیت بیشتر CPU، حافظه یا I/O دیسک به دلیل اجرای اسکریپت‌های دستکاری خودکار در پس‌زمینه.

- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول کاربران ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به احراز هویت ناموفق، فرآیندهای غیرمعمول یا فعالیت غیرمنتظره.
- **ترافیک شبکه غیرمنتظره:**
  - انحرافات در ترافیک خروجی، مانند اتصالات به آدرس‌های IP یا دامنه‌های ناشناخته، که ممکن است نشان‌دهنده خروج داده‌های تغییر یافته باشد.

۷-۴۹-۲ تأثیرات

- **از دست دادن یکپارچگی داده:**
  - فساد رکوردهای حیاتی منجر به بی‌اعتمادی به سیستم‌ها، تصمیم‌گیری نادرست یا شکست عملیاتی می‌شود.
- **ضررهای مالی:**
  - تراکنش‌های غیرمجاز، فعالیت‌های کلاهبرداری یا اختلال در خدمات که باعث آسیب‌های مالی می‌شوند.
- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به‌دلیل رکوردهای به‌خطرافتاده یا لاگ‌های دستکاری‌شده.
- **آسیب به شهرت:**
  - از دست دادن اعتماد مشتریان و ارزش برند پس از یک نفوذ یا سوءاستفاده از داده‌های دستکاری‌شده.
- **جریمه‌های نظارتی:**
  - عدم رعایت مقررات حفاظت از داده‌ها، که منجر به جریمه‌ها و اقدامات قانونی می‌شود.
- **خطرات ایمنی:**
  - در بهداشت یا زیرساخت‌های حیاتی، رکوردهای دستکاری‌شده می‌توانند جان افراد را به خطر بیندازند یا باعث شکست‌های فاجعه‌بار شوند.

۸-۴۹-۲ راه‌های مقابله

- **کنترل‌های دسترسی:**
  - کنترل‌های دسترسی سخت‌گیرانه، احراز هویت چندعاملی و اصل حداقل دسترسی را اعمال کنید تا دسترسی غیرمجاز را محدود کنید.
- **بررسی‌های یکپارچگی داده:**
  - مکانیزم‌هایی مانند هش‌کردن، امضای دیجیتال یا تأیید مبتنی بر بلاک‌چین را پیاده‌سازی کنید تا اصالت و یکپارچگی رکوردها را تضمین کنید.
- **حسابرسی‌های منظم:**
  - حسابرسی‌های مکرر از دسترسی و تغییرات داده‌ها انجام دهید تا فعالیت‌های مشکوک را به‌طور پیش‌گیرانه شناسایی و برطرف کنید.
- **نظارت رفتاری:**
  - از تحلیل رفتاری مبتنی بر هوش مصنوعی استفاده کنید تا الگوهای فعالیت غیرعادی که نشان‌دهنده دستکاری رکوردها هستند، شناسایی شوند.
- **رمزنگاری:**
  - داده‌های حساس را هم در حالت ذخیره‌سازی و هم در حال انتقال رمزنگاری کنید تا ارزش اطلاعات سرقت‌شده یا دستکاری‌شده را به حداقل برسانید.
- **ردیابی حسابرسی:**
  - ردهای حسابرسی دقیق و تغییرناپذیر را با استفاده از فناوری‌هایی مانند بلاک‌چین حفظ کنید تا تمام تغییرات داده‌ها را ردیابی کنید.
- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد شناسایی نشانه‌های دستکاری و رعایت بهداشت سایبری آموزش دهید.
- **هوشمندسازی تهدیدات پیشرفته:**
  - از پلتفرم‌های هوشمند مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور دستکاری و تهدیدات تطبیقی مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی‌ای توسعه دهید و پیاده‌سازی کنید تا حملات دستکاری را

به‌سرعت شناسایی، مهار و کاهش دهید.

- **دفاع پیش‌گیرانه:**
  - از یادگیری ماشین تقابلی استفاده کنید تا اقدامات دفاعی را در برابر استراتژی‌های در حال تکامل دستکاری شبیه‌سازی و تست کنید.

۲-۵۰ محروم‌سازی از خدمت DDos

حمله انکار خدمات توزیع‌شده (DDoS)[1] شامل ازدحام بیش از حد یک سیستم، سرویس یا شبکه هدف با ترافیک زیاد است که باعث می‌شود سیستم برای کاربران قانونی غیرقابل دسترس شود. در حمله DDoS مبتنی بر هوش مصنوعی، از هوش مصنوعی برای افزایش دقت، کارایی و انعطاف‌پذیری تکنیک‌های سنتی DDoS استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند مدیریت بات‌نت‌ها را بهینه‌سازی کنند، بردارهای حمله را به صورت پویا تنظیم کنند و از اقدامات دفاعی مانند محدود کردن نرخ ترافیک، فیلتر کردن ترافیک یا تشخیص ناهنجاری‌ها فرار کنند.

این نوع حمله می‌تواند از دفاع‌های سنتی عبور کند، به سرعت مقیاس‌پذیری داشته باشد و به صورت بلادرنگ به شرایط متغیر واکنش نشان دهد، که این امر باعث اختلالات قابل توجه در سرویس‌های حیاتی می‌شود.

۲-۵۰-۱ ویژگی‌های کلیدی

- **خودکار سازی:**
  - هوش مصنوعی فرآیند شناسایی اهداف، جذب بات‌ها و راه‌اندازی حملات را خودکار می‌کند، که باعث کاهش تلاش دستی و افزایش سرعت می‌شود.
- **مقیاس‌پذیری:**
  - الگوریتم‌های یادگیری ماشین به مهاجمان امکان می‌دهند تا بات‌نت‌های بزرگ‌مقیاس متشکل از هزاران یا میلیون‌ها دستگاه آلوده را هماهنگ کنند.
- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی می‌توانند با گذشت زمان تکامل یابند و با تغییرات در پیکربندی شبکه، مکانیزم‌های دفاعی یا عوامل محیطی سازگار شوند.
- **دقت:**
  - الگوریتم‌های پیشرفته به مهاجمان اجازه می‌دهند تا آسیب‌پذیری‌ها یا نقاط ضعف خاص در

1. Distributed Denial of Service (DDOS)

زیرساخت هدف را شناسایی کنند تا بیشترین تأثیر را داشته باشند.

- **پنهان‌کاری:**
  - با تقلید از الگوهای ترافیک قانونی یا استفاده از تکنیک‌های پیچیده پنهان‌سازی، حملات DDoS مبتنی بر هوش مصنوعی می‌توانند برای مدت طولانی‌تری شناسایی نشوند

۲-۵۰-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های برچسب‌دار از حملات DDoS موفق و ناموفق آموزش می‌دهد تا استراتژی‌های حمله را بهبود بخشد.
  - یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها در ترافیک شبکه را شناسایی می‌کند تا نقاط ضعف یا بردارهای حمله بهینه را کشف کند.
  - یادگیری تقویتی: پارامترهای حمله را با پاداش دادن به نتایج موفق و جریمه کردن شکست‌ها بهینه‌سازی می‌کند، که امکان بهبود مستمر را فراهم می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در ترافیک شبکه را تحلیل می‌کنند تا استراتژی‌های حمله مؤثر را پیش‌بینی کنند.
  - شبکه‌های مولد تخاصمی: الگوهای ترافیک مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا تکنیک‌های فرار را آزمایش و بهبود بخشند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی کاربر یا سیستم استفاده می‌کند، که به مهاجمان امکان می‌دهد ترافیکی ایجاد کنند که به‌راحتی با فعالیت قانونی ترکیب شود.
- **پنهان‌سازی ترافیک:**
  - از هوش مصنوعی برای رمزگذاری، قطعه‌قطعه کردن یا اضافه کردن داده‌های اضافی به ترافیک مخرب استفاده می‌کند، که شناسایی یا مسدود کردن آن را برای سیستم‌های دفاعی سخت‌تر می‌کند.
- **مدیریت بات‌نت:**
  - از هوش مصنوعی برای جذب، سازماندهی و مدیریت کارآمد بات‌نت‌های بزرگ‌مقیاس استفاده می‌کند، که باعث استفاده بهینه از منابع در طول حمله می‌شود.

۳-۵۰-۲ دارایی‌های هدف

- **برنامه‌های وب:**
  - پلتفرم‌های تجارت الکترونیک، مؤسسات مالی یا وب‌سایت‌های دولتی که در برابر اختلال سرویس آسیب‌پذیر هستند.
- **سرویس‌های ابری:**
  - برنامه‌های میزبانی‌شده در ابر، APIها یا سیستم‌های ذخیره‌سازی که در معرض اینترنت قرار دارند و حملات DDoS می‌توانند باعث اختلال گسترده در آن‌ها شوند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تأسیسات بهداشتی، شبکه‌های حمل‌ونقل یا سیستم‌های ارتباطی که downtime در آن‌ها می‌تواند عواقب شدیدی داشته باشد.
- **دستگاه‌های اینترنت اشیا (IoT):**
  - دستگاه‌های IoT با منابع محدود که اغلب به دلیل ضعف در اقدامات امنیتی به عنوان بخشی از بات‌نت‌ها استفاده می‌شوند.
- **شبکه‌های سازمانی:**
  - شبکه‌های شرکتی که داده‌های حساس را ذخیره می‌کنند یا میزبان سرویس‌های حیاتی هستند.

۴-۵۰-۲ آسیب‌پذیری‌ها

- **امنیت ضعیف بات‌نت:**
  - دستگاه‌های IoT، روترها یا سایر سیستم‌های متصل که دارای اعتبارنامه‌های پیش‌فرض یا ریزبرنامه قدیمی هستند، به راحتی به بات‌نت‌ها اضافه می‌شوند.
- **فیلتر کردن ناکافی ترافیک:**
  - عدم وجود مکانیزم‌های قوی فیلتر کردن ترافیک باعث می‌شود ترافیک مخرب سیستم‌های هدف را تحت تأثیر قرار دهد.
- **خطای انسانی:**
  - پیکربندی نادرست فایروال‌ها، متعادل‌کننده‌های بار یا سرورهای DNS به طور ناخواسته آسیب‌پذیری‌هایی را ایجاد می‌کند که توسط حملات DDoS قابل بهره‌برداری است.

- **دفاع‌های قدیمی:**
  - استفاده از راه‌حل‌های قدیمی یا نادرست برای کاهش DDoS، موفقیت حملات مبتنی بر هوش مصنوعی را آسان‌تر می‌کند.
- **سواستفاده از آسیب‌پذیری روز صفر:**
  - آسیب‌پذیری‌های کشف‌نشده در نرم‌افزار یا سخت‌افزار، نقاط ورودی را برای مهاجمان فراهم می‌کنند تا بات‌نت‌ها را ایجاد کنند یا حملات را تقویت کنند.

۵-۵۰-۲ سناریوی تهدید

- **شناسایی:**
  - از ابزارهای مبتنی بر هوش مصنوعی برای نقشه‌برداری از شبکه هدف، شناسایی گلوگاه‌ها و ارزیابی ظرفیت آن برای مدیریت حجم بالای ترافیک استفاده می‌شود.
- **جذب بات‌نت:**
  - الگوریتم‌های یادگیری ماشین برای اسکن دستگاه‌های آسیب‌پذیر به کار گرفته می‌شوند تا با بهره‌برداری از ضعف‌ها، آن‌ها را به بات‌نت اضافه کنند.
- **برنامه‌ریزی حمله:**
  - داده‌های تاریخی تحلیل و سناریوهای حمله شبیه‌سازی می‌شوند تا مؤثرترین ترکیب از بردارهای حمله، زمان‌بندی و شدت تعیین شود.
- **تولید ترافیک:**
  - بات‌نت هماهنگ می‌شود تا حجم عظیمی از ترافیک را ایجاد کند و منابع یا سرویس‌های خاص در زیرساخت قربانی را هدف قرار دهد.
- **سازگاری:**
  - اثربخشی حمله به طور مداوم نظارت می‌شود و پارامترها به صورت بلادرنگ تنظیم می‌شوند تا از اقدامات دفاعی مانند محدود کردن نرخ ترافیک یا فیلتر کردن عبور کنند.
- **پایداری:**
  - حمله برای مدت طولانی حفظ می‌شود و از بردارها یا تکنیک‌های تقویت مختلف استفاده می‌کند تا فشار بر هدف ادامه یابد.

۲-۵۰-۶ نشانه‌های احتمال نفوذ

- **افزایش حجم ترافیک:**
  - ناهنجاری‌ها در ترافیک ورودی یا خروجی، مانند افزایش ناگهانی یا الگوهای غیرعادی.
- **کاهش کیفیت سرویس:**
  - زمان پاسخ‌دهی کندتر، افزایش تأخیر یا عدم دسترسی کامل به سرویس‌های هدف.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به اتصالات ناموفق، فعالیت‌های غیرمنتظره یا تلاش‌های دسترسی غیرمجاز.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار معمول شبکه، مانند درخواست‌های مکرر از آدرس‌های IP یا مناطق جغرافیایی ناشناس.
- **اتمام منابع:**
  - استفاده بیشتر از CPU، حافظه یا پهنای باند که با عملیات قانونی مرتبط نیست.

۲-۵۰-۷ تأثیرات

- **اختلال عملیاتی:**
  - وقفه در سرویس‌ها یا سیستم‌های حیاتی کسب‌وکار، که منجر به از دست دادن درآمد و آسیب به اعتبار می‌شود.
- **ضررهای مالی:**
  - هزینه‌های مرتبط با downtime، پاسخ اضطراری یا جبران خسارت به مشتریان آسیب‌دیده.
- **آسیب به اعتبار:**
  - از دست دادن اعتماد مشتریان و ارزش برند پس از اختلالات طولانی‌مدت یا کاهش کیفیت سرویس.
- **جریمه‌های قانونی:**
  - عدم رعایت توافق‌نامه‌های سطح سرویس (SLAs) یا الزامات قانونی به دلیل عدم دسترسی طولانی‌مدت.
- **حملات ثانویه:**
  - حملات DDoS ممکن است به عنوان حواس‌پرتی برای نفوذهای شدیدتر، مانند نقض داده‌ها

یا عفونت‌های باج‌افزار، استفاده شوند.

۸-۵۰-۲ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های IDS/IPS مبتنی بر هوش مصنوعی را مستقر کنید که قادر به تشخیص الگوهای ترافیک غیرعادی مرتبط با حملات DDoS هستند.
- **فیلتر کردن ترافیک:**
  - مکانیزم‌های پیشرفته فیلتر کردن ترافیک، مانند محدود کردن نرخ، blackholing یا sinkhol-ing را پیاده‌سازی کنید تا اثرات ترافیک بیش از حد را کاهش دهید.
- **توزیع بار:**
  - از متعادل‌کننده‌های بار مبتنی بر هوش مصنوعی استفاده کنید تا ترافیک را بین چندین سرور توزیع کند و خطر خرابی تک‌نقطه‌ای را کاهش دهد.
- **ادغام با شبکه‌های تحویل محتوا (CDN)[1]:**
  - از شبکه‌های تحویل محتوا استفاده کنید تا ترافیک مخرب را قبل از رسیدن به سیستم هدف جذب و فیلتر کند.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی استفاده کنید تا انحرافات ظریف در الگوهای ترافیک را تشخیص داده و به آن‌ها پاسخ دهید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، ریزبرنامه و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد استفاده بات‌نت‌های DDoS را برطرف کنید.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش ترافیک مخرب را محدود کرده و تأثیر حملات موفق را کاهش دهید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به سرعت حملات DDoS را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.

1. Content Delivery Network (CDN)

- **هوشمندی تهدیدات پیش‌فعال:**
  - از پلتفرم‌های هوشمندی تهدیدات مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور DDoS و تهدیدات انطباقی مطلع شوید.
- **آموزش و آگاهی:**
  - به کارمندان و کاربران آموزش دهید تا علائم حملات DDoS، مانند عملکرد کند یا سرویس‌های غیرقابل دسترس، را تشخیص دهند.

۲-۵۱ باج‌افزارهای مبتنی بر هوش مصنوعی

حملات باج‌افزاری[1] مبتنی بر هوش مصنوعی نشان‌دهنده شکل جدید و بسیار پیشرفته‌ای از جرایم سایبری هستند که در آن از هوش مصنوعی برای افزایش دقت، انعطاف‌پذیری و اثربخشی باج‌افزارهای سنتی استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند فرآیند شناسایی اهداف باارزش، فرار از سیستم‌های تشخیص و بهینه‌سازی تکنیک‌های رمزنگاری را خودکار کنند. این نوع حمله به‌ویژه خطرناک است زیرا می‌تواند از اقدامات امنیتی پیشرفته عبور کند، به‌طور لحظه‌ای با استراتژی‌های دفاعی سازگار شود و تأثیر خود را بر قربانیان به حداکثر برساند.

باج‌افزارها مدت‌هاست که به عنوان یک تهدید جدی مطرح هستند، اما انواع مبتنی بر هوش مصنوعی این خطر را به سطح جدیدی می‌برند، زیرا به مهاجمان امکان می‌دهند عملیات خود را گسترش دهند، زیرساخت‌های حیاتی را هدف قرار دهند و با نرخ موفقیت بالاتری باج‌های بیشتری طلب کنند.

۲-۵۱-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی کل چرخه حیات حملات باج‌افزاری، از شناسایی تا رمزنگاری و مذاکره برای دریافت باج، را خودکار می‌کند.
- **انعطاف‌پذیری:**
  - باج‌افزارهای مبتنی بر هوش مصنوعی به‌طور پویا تکامل می‌یابند و با تغییرات در دفاع‌های سیستم، پیکربندی شبکه یا رفتار کاربر سازگار می‌شوند.
- **دقت هدف‌گیری:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا اهداف باارزش، مانند سیستم‌های

1. Ransomware Attacks

حیاتی، داده‌های حساس یا نقاط پایانی آسیب‌پذیر را شناسایی کنند.

- **پنهان‌کاری:**
  - با تقلید از فرآیندهای قانونی یا رمزنگاری payloadها تا زمان اجرا، هوش مصنوعی اطمینان می‌دهد که باج‌افزار برای مدت طولانی بدون تشخیص باقی بماند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد باج‌افزار را به‌طور همزمان در چندین سیستم یا شبکه مستقر کنند، که دامنه و تأثیر حمله را افزایش می‌دهد.

۲-۵۱-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های موفق و ناموفق حملات باج‌افزاری آموزش می‌دهد تا استراتژی‌های هدف‌گیری و فرار را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا ناهنجاری‌ها را در رفتار سیستم شناسایی می‌کند تا آسیب‌پذیری‌ها یا مکانیسم‌های دفاعی بالقوه را آشکار کند.
  - یادگیری تقویتی: تاکتیک‌های باج‌افزاری را با پاداش دادن به نتایج موفق (مانند انتشار بدون تشخیص) و جریمه کردن شکست‌ها بهینه می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی پیچشی و شبکه‌های عصبی بازگشتی، الگوهای پیچیده در ترافیک شبکه، ساختار فایل‌ها یا تعاملات کاربر را تحلیل می‌کنند تا قابلیت‌های فرار و هدف‌گیری را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: payloadها یا رفتارهای مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های تطبیقی را آزمایش و بهبود بخشند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار طبیعی کاربر یا سیستم استفاده می‌کند تا مهاجمان بتوانند انحرافات را تشخیص دهند یا موجودیت‌های قانونی را تقلید کنند.
- **تشخیص محیط:**
  - از هوش مصنوعی برای تشخیص محیط‌های sandbox ، تله‌های هانی‌پات یا سایر اقدامات دفاعی استفاده می‌کند و رفتار را بر این اساس تغییر می‌دهد.

- **پردازش زبان طبیعی:**
  - داده‌های متنی، مانند پیام‌های خطا، لاگ‌ها یا ارتباطات را تحلیل می‌کند تا اطلاعات بیشتری درباره سیستم هدف استنباط کند.

۲-۵۱-۳ دارایی‌های هدف

- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، مراکز بهداشتی، مؤسسات مالی یا شبکه‌های حمل‌ونقل که downtime می‌تواند آسیب‌های جدی ایجاد کند.
- **سیستم‌های سازمانی:**
  - شبکه‌های شرکتی که مالکیت فکری، اسرار تجاری یا داده‌های مشتریان ارزشمند را ذخیره می‌کنند.
- **سازمان‌های دولتی:**
  - اهداف باارزش که اطلاعات طبقه‌بندی‌شده را ذخیره می‌کنند یا سیستم‌های امنیت ملی را کنترل می‌کنند.
- **کسب‌وکارهای کوچک و متوسط (SMBs):**
  - سازمان‌هایی که منابع امنیت سایبری محدودی دارند و آن‌ها را به اهدافی آسان برای باج‌افزارهای مبتنی بر هوش مصنوعی تبدیل می‌کند.
- **افراد:**
  - دستگاه‌های شخصی یا حساب‌هایی که حاوی داده‌های حساس، مانند عکس‌ها، اسناد یا سوابق مالی هستند.

۲-۵۱-۴ آسیب‌پذیری‌ها

- **امنیت ضعیف نقاط پایانی:**
  - نبود راه‌حل‌های ضدویروس قوی، فایروال‌ها یا ابزارهای محافظت از نقاط پایانی، سیستم‌ها را در برابر عفونت‌های باج‌افزاری آسیب‌پذیر می‌کند.
- **خطای انسانی:**
  - کارمندان یا کاربرانی که در دام ایمیل‌های فیشینگ، دانلودهای مخرب یا تاکتیک‌های مهندسی اجتماعی می‌افتند، به‌طور ناخواسته سیستم‌های خود را در معرض خطر قرار می‌دهند.

- **سیستم‌های قدیمی:**
  - استفاده از نرم‌افزارها، فرم‌ورها یا پروتکل‌های قدیمی که حاوی آسیب‌پذیری‌های شناخته‌شده هستند و توسط باج‌افزارها قابل سوءاستفاده هستند.
- **پشتیبان‌گیری ناکافی:**
  - روش‌های ضعیف پشتیبان‌گیری سازمان‌ها را قادر نمی‌سازد بدون پرداخت باج از حملات موفق بهبود یابند.
- **عدم وجود اقدامات متقابل:**
  - نبود دفاع‌های مبتنی بر هوش مصنوعی یا هوش تهدید پیش‌گیرانه، سیستم‌ها را در برابر انواع در حال تحول باج‌افزارها آسیب‌پذیر می‌کند.

۵-۵۱-۲ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره هدف، از جمله معماری، پیکربندی‌ها و رفتارهای معمول آن استفاده می‌کند.
- **اسکن آسیب‌پذیری:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل هدف از نظر نقاط ضعف، مانند آسیب‌پذیری‌های اصلاح‌نشده یا مکانیسم‌های احراز هویت ضعیف، مستقر می‌کند.
- **تولید payload:**
  - از یادگیری عمیق یا شبکه‌های مولد تخاصمی برای تولید payloadهای سفارشی‌سازی‌شده که متناسب با آسیب‌پذیری‌های شناسایی‌شده هستند استفاده می‌کند.
- **استقرار:**
  - مهاجم باج‌افزار را از طریق ایمیل‌های فیشینگ، کیت‌های اکسپلویت یا سایر بردارها تحویل می‌دهد و اطمینان حاصل می‌کند که با ترافیک عادی ترکیب می‌شود.
- **اجرای رمزنگاری:**
  - مهاجم روال‌های رمزنگاری را اجرا کرده تا فایل‌ها، سیستم‌ها یا پایگاه‌داده‌های حیاتی را قفل کرده و در عین حال بدون تشخیص باقی بماند.
- **درخواست باج:**
  - مهاجم از پردازش زبان طبیعی برای ایجاد یادداشت‌های باج شخصی‌سازی‌شده استفاده

می‌کند تا فشار روانی بر قربانیان افزایش یابد.

- **پایداری:**
  - مهاجم درهای پشتی ایجاد می‌کند یا نقاط دسترسی را حفظ کرده تا حملات مکرر یا استخراج داده تسهیل شود.

۲-۵۱-۶ نشانه‌های احتمال نفوذ

- **تغییرات غیرمعمول فایل:**
  - ناهنجاری‌ها در الگوهای ایجاد، حذف یا تغییر فایل‌ها که با فعالیت‌های قانونی مرتبط نیستند.
- **افزایش استفاده از منابع:**
  - فعالیت بیشتر CPU، حافظه یا I/O دیسک به دلیل فرآیندهای رمزنگاری که در پس‌زمینه اجرا می‌شوند.
- **شاخص‌های رفتاری:**
  - اقداماتی که با رفتار معمول سیستم ناسازگار هستند، مانند دسترسی به مناطق محدود یا انجام پرس‌وجوهای غیرمعمول.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.
- **ترافیک شبکه غیرمنتظره:**
  - انحراف در ترافیک خروجی، مانند اتصالات به آدرس‌های IP یا دامنه‌های ناشناخته، که ممکن است نشان‌دهنده ارتباطات فرمان و کنترل (C2) باشد.

۲-۵۱-۷ تأثیرات

- **اختلال عملیاتی:**
  - رمزنگاری فایل‌ها یا سیستم‌های حیاتی منجر به downtime و اختلال عملیاتی قابل توجه می‌شود.
- **ضررهای مالی:**
  - قربانیان ممکن است با درخواست‌های اخاذی به شکل پرداخت‌های ارز دیجیتال مواجه شوند که منجر به ضررهای مالی مستقیم می‌شود.

- **نقض داده‌ها:**
  - برخی از انواع باج‌افزارهای مبتنی بر هوش مصنوعی قبل از رمزنگاری داده‌ها را سرقت می‌کنند، که منجر به نقض داده‌های اضافی و مسائل مربوط به انطباق می‌شود.
- **آسیب به شهرت:**
  - سازمان‌هایی که از حملات باج‌افزاری رنج می‌برند ممکن است اعتماد مشتریان و ارزش برند خود را از دست بدهند.
- **جریمه‌های نظارتی:**
  - عدم انطباق با مقررات حفاظت از داده‌ها، ممکن است منجر به جریمه‌ها و اقدامات قانونی شود.

۸-۵۱-۲ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - راه‌حل‌های ضدویروس مبتنی بر هوش مصنوعی را مستقر کنید که بر روی مجموعه‌داده‌های بزرگی از نمونه‌های باج‌افزاری آموزش دیده‌اند تا تهدیدهای مبهم را به‌طور مؤثرتری شناسایی کنند.
- **تحلیل رفتاری:**
  - از تحلیل‌های رفتاری مبتنی بر هوش مصنوعی برای تشخیص فعالیت‌های غیرعادی که نشان‌دهنده باج‌افزار هستند استفاده کنید، حتی اگر امضای شناخته‌شده‌ای نداشته باشند.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فرم‌ورها و وابستگی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده توسط باج‌افزارها اصلاح شوند.
- **استراتژی‌های پشتیبان‌گیری قوی:**
  - پشتیبان‌گیری‌های منظم، ایمن و آفلاین را پیاده‌سازی کنید تا بازیابی داده‌ها بدون پرداخت باج تضمین شود.
- **تقسیم‌بندی شبکه:**
  - شبکه را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش باج‌افزار محدود شود و تأثیر عفونت‌های موفق کاهش یابد.

- **آموزش و آگاهی:**
  - کارمندان و کاربران را در مورد تشخیص علائم باج‌افزار، مانند تلاش‌های فیشینگ یا رفتار غیرمعمول سیستم آموزش دهید.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا درباره تکنیک‌های نوظهور باج‌افزاری و تهدیدات تطبیقی به‌روز بمانید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حملات باج‌افزاری را تشخیص، مهار و کاهش دهید.
- **معماری Zero-Trust:**
  - رویکرد zero-trust را برای امنیت شبکه اتخاذ کنید، که هر تلاش دسترسی را بدون توجه به منشأ آن تأیید می‌کند.
- **دفاع پیش‌گیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر انواع در حال تحول باج‌افزارها استفاده کنید.

۲-۵۲ دستکاری افکار عمومی

دستکاری افکار عمومی[1] با استفاده از هوش مصنوعی به استفاده از فناوری‌های هوش مصنوعی برای تأثیرگذاری، شکل‌دهی یا تحریف ادراک عمومی در مقیاس بزرگ اشاره دارد. با بهره‌گیری از یادگیری ماشین، پردازش زبان طبیعی و یادگیری عمیق، مهاجمان می‌توانند حجم عظیمی از داده‌ها را از شبکه‌های اجتماعی، رسانه‌های خبری و سایر منابع تحلیل کنند تا روایت‌های بسیار متقاعدکننده‌ای ایجاد کنند، اطلاعات نادرست را گسترش دهند یا محتوای تفرقه‌انگیز را تقویت کنند. این نوع دستکاری به‌ویژه خطرناک است زیرا اعتماد به فرآیندهای دموکراتیک، نهادهای رسانه‌ای و انسجام اجتماعی را تضعیف می‌کند.

هوش مصنوعی به مهاجمان امکان می‌دهد تا ایجاد و انتشار محتوا را خودکار کنند، جمعیت‌های خاص را با دقت هدف قرار دهند و کمپین‌ها را بر اساس بازخوردهای لحظه‌ای تطبیق دهند. در نتیجه، دستکاری افکار عمومی با استفاده از هوش مصنوعی خطرات قابل توجهی برای افراد، سازمان‌ها و کل جوامع به همراه دارد.

1. Public Opinion Manipulation

۲-۵۲-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند تولید، توزیع و تقویت محتوا را خودکار می‌کند، که باعث کاهش تلاش دستی و افزایش کارایی می‌شود.
- **هدف‌گیری دقیق:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند تا پیام‌ها را بر اساس ترجیحات، رفتارها یا آسیب‌پذیری‌های مخاطبان خاص تنظیم کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی امکان تولید و توزیع سریع حجم زیادی از محتوا را در چندین پلتفرم به‌طور همزمان فراهم می‌کند.
- **انعطاف‌پذیری:**
  - کمپین‌ها می‌توانند با گذشت زمان تکامل یابند و با تغییرات در احساسات عمومی، اقدامات دفاعی یا شرایط محیطی سازگار شوند.
- **پنهان‌کاری:**
  - با تقلید از سبک‌های ارتباطی قانونی یا سوءاستفاده از سوگیری‌های الگوریتمی، هوش مصنوعی اطمینان می‌دهد که محتوای دستکاری‌شده بدون تشخیص باقی بماند.

۲-۵۲-۳ تکنیک‌های هوش مصنوعی مورد استفاده

- **پردازش زبان طبیعی:**
  - تولید متن: از مدل‌های مبتنی بر ترنسفورمر مانند GPT-3 یا BERT برای تولید مقالات، پست‌ها یا نظرات منسجم و مرتبط با زمینه استفاده می‌کند.
  - تحلیل احساسات: احساسات عمومی را تحلیل می‌کند تا موضوعات یا تم‌هایی را که با مخاطبان هدف طنین‌انداز می‌شوند شناسایی کند.
  - مدل‌سازی موضوعی: موضوعات داغ یا جنجال‌ها را شناسایی می‌کند تا روایت‌های دستکاری‌شده حول آن‌ها ایجاد کند.
- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را بر روی مجموعه‌داده‌های موفق و ناموفق دستکاری آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.

- یادگیری بدون نظارت: خوشه‌ها یا ناهنجاری‌ها را در داده‌های ارتباطی شناسایی می‌کند تا نقاط ضعف یا فرصت‌های بالقوه را آشکار کند.
- یادگیری تقویتی: کمپین‌ها را با پاداش دادن به معیارهای تعامل موفق (مانند لایک‌ها، اشتراک‌ها، بازدیدها) و جریمه کردن شکست‌ها بهینه می‌کند.

- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و شبکه‌های مولد تخاصمی، محتوای چندرسانه‌ای مصنوعی اما واقع‌گرایانه مانند جعل عمیق‌ها، تصاویر دستکاری‌شده یا ویدیوهای جعلی ایجاد می‌کنند.
- **تحلیل شبکه‌های اجتماعی:**
  - شبکه‌های عصبی گرافی ساختار شبکه‌های اجتماعی را تحلیل می‌کنند تا گره‌ها یا جوامع تأثیرگذار را برای انتشار هدفمند شناسایی کنند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار طبیعی کاربر استفاده می‌کند تا به مهاجمان امکان تقلید از فعالیت‌های قانونی و فرار از تشخیص را بدهد.

۳-۵۲-۲ دارایی‌های هدف

- **افکار عمومی:**
  - باورها، نگرش‌ها و ادراک افراد یا گروه‌ها درباره مسائل سیاسی، اجتماعی یا اقتصادی.
- **انتخابات و حکمرانی:**
  - فرآیندهای دموکراتیک، جایی که دستکاری می‌تواند نظرات رأی‌دهندگان را تحت تأثیر قرار دهد، نتایج انتخابات را تغییر دهد یا دولت‌ها را بی‌ثبات کند.
- **پلتفرم‌های رسانه‌ای:**
  - شبکه‌های اجتماعی، وب‌سایت‌های خبری یا فروم‌هایی که محتوای دستکاری‌شده می‌تواند از طریق توصیه‌های الگوریتمی یا اشتراک‌گذاری ویروسی تقویت شود.
- **بازارهای مالی:**
  - قیمت سهام، تصمیمات سرمایه‌گذاری یا ثبات بازار که تحت تأثیر اطلاعات نادرست درباره شرکت‌ها، صنایع یا شاخص‌های اقتصادی قرار می‌گیرند.
- **زیرساخت‌های حیاتی:**
  - اعتماد به خدمات ضروری، مانند مراقبت‌های بهداشتی، انرژی یا حمل‌ونقل، که با انتشار

ترس، عدم اطمینان یا تردید تضعیف می‌شود.

۲-۵۲-۴ آسیب‌پذیری‌ها

- **روانشناسی انسان:**
  - افراد بیشتر احتمال دارد که محتوای احساسی یا جنجالی را باور کنند، که آن‌ها را در برابر دستکاری‌های تولیدشده توسط هوش مصنوعی آسیب‌پذیر می‌کند.
- **تقویت الگوریتمی:**
  - الگوریتم‌های شبکه‌های اجتماعی محتوای جذاب را در اولویت قرار می‌دهند و به‌طور ناخواسته اطلاعات دستکاری‌شده یا گمراه‌کننده را ترویج می‌کنند.
- **مکانیسم‌های ضعیف بررسی واقعیت:**
  - منابع محدود یا تکنیک‌های قدیمی برای تأیید اعتبار محتوای آنلاین باعث می‌شود دستکاری بدون کنترل گسترش یابد.
- **نشت داده‌ها:**
  - مجموعه‌داده‌های نشت‌شده حاوی اطلاعات شخصی مواد ارزشمندی برای ایجاد کمپین‌های دستکاری هدفمند فراهم می‌کنند.
- **سوءاستفاده از پلتفرم‌ها:**
  - نبود ابزارها یا سیاست‌های نظارتی قوی در برخی پلتفرم‌ها به ربات‌های خودکار و ترول‌ها امکان می‌دهد محتوای دستکاری‌شده را آزادانه منتشر کنند.

۲-۵۲-۵ سناریوی تهدید

- **جمع‌آوری داده:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره مخاطبان هدف، از جمله ترجیحات، رفتارها و آسیب‌پذیری‌های آن‌ها استفاده می‌کند.
- **تولید محتوا:**
  - مهاجم از مدل‌های پردازش زبان طبیعی برای ایجاد محتوای متقاعدکننده و متناسب با زمینه که مختص گروه‌ها یا افراد خاص است استفاده می‌کند.
- **ایجاد محتوای چندرسانه‌ای:**
  - مهاجم از شبکه‌های مولد تخاصمی یا سایر تکنیک‌های یادگیری عمیق برای تولید تصاویر، ویدیوها یا کلیپ‌های صوتی مصنوعی که از روایت پشتیبانی می‌کنند استفاده می‌کند.

- **انتشار:**
  - محتوای دستکاری‌شده در چندین پلتفرم با استفاده از بات‌نت‌ها یا حساب‌های نفوذشده به طور خودکار ارسال می‌شود.
- **تقویت:**
  - مهاجم از الگوریتم‌های شبکه‌های اجتماعی، موضوعات داغ یا شبکه‌های تأثیرگذار برای حداکثر رساندن دسترسی و تأثیر استفاده می‌کند.
- **پایداری:**
  - مهاجم کمپین را به‌طور مداوم بر اساس بازخوردهای لحظه‌ای بهبود می‌بخشد تا تأثیر بلندمدت حفظ شود.

۲-۵۲-۶ نشانه‌های احتمال نفوذ

- **الگوهای غیرمعمول محتوا:**
  - ناهنجاری‌ها در فرکانس ارسال، لحن یا سبک که از رفتار معمول کاربر منحرف می‌شوند.
- **فعالیت ربات‌ها:**
  - حجم بالایی از محتوای یکسان یا مشابه که توسط چندین حساب در بازه زمانی کوتاهی ارسال می‌شود.
- **معیارهای تعامل:**
  - افزایش ناگهانی لایک‌ها، اشتراک‌ها یا نظرات روی محتوای جنجالی یا احساسی.
- **حساب‌های غیرواقعی:**
  - پروفایل‌های مشکوک با جزئیات ناقص، تاریخ‌های ایجاد اخیر یا سطح فعالیت غیرعادی.
- **محتوای چندرسانه‌ای گمراه‌کننده:**
  - تصاویر، ویدیوها یا کلیپ‌های صوتی تغییر یافته یا جعلی که با حقایق تأییدشده در تضاد هستند.

۲-۵۲-۷ تأثیرات

- **تفرقه اجتماعی:**
  - قطبی‌شدن افکار عمومی، فرسایش اعتماد به نهادها و افزایش ناآرامی‌های اجتماعی.
- **پیامدهای اقتصادی:**
  - نوسانات بازار، ضررهای مالی یا آسیب‌های اعتباری به کسب‌وکارها به دلیل اطلاعات نادرست.

- **تأثیرات سیاسی:**
  - دستکاری نتایج انتخابات، ایجاد اختلاف میان رأی‌دهندگان یا دخالت در روابط بین‌المللی.
- **خطرات بهداشت عمومی:**
  - انتشار اطلاعات نادرست درباره بحران‌های بهداشتی، واکسن‌ها یا درمان‌های پزشکی که منجر به رفتارهای مضر می‌شود.
- **چالش‌های نظارتی:**
  - مشکل در اجرای قوانین یا مقررات علیه اشکال در حال تحول دستکاری.

۸-۵۲-۲ راه‌های مقابله

- **تشخیص پیشرفته محتوا:**
  - از ابزارهای مبتنی بر هوش مصنوعی استفاده کنید که قادر به شناسایی محتوای مصنوعی یا دستکاری‌شده، مانند تشخیص‌دهنده‌های جعل عمیق یا بررسی‌کننده‌های سرقت ادبی هستند.
- **نظارت رفتاری:**
  - از هوش مصنوعی برای تشخیص الگوهای غیرمعمول فعالیت که نشان‌دهنده بات‌نت‌ها یا کمپین‌های دستکاری هماهنگ هستند استفاده کنید.
- **خودکارسازی بررسی واقعیت:**
  - سیستم‌های بررسی واقعیت خودکار را توسعه دهید که با استفاده از پردازش زبان طبیعی قادر به تأیید صحت محتوای آنلاین در زمان واقعی هستند.
- **نظارت بر پلتفرم‌ها:**
  - سیاست‌ها و ابزارهای نظارتی سخت‌گیرانه‌تری را در شبکه‌های اجتماعی پیاده‌سازی کنید تا از گسترش محتوای دستکاری‌شده جلوگیری شود.
- **آموزش کاربران:**
  - کاربران را آموزش دهید تا محتوای آنلاین را به‌طور انتقادی ارزیابی کنند، علائم دستکاری را تشخیص دهند و فعالیت‌های مشکوک را گزارش دهند.
- **تلاش‌های مشترک:**
  - همکاری بین دولت‌ها، شرکت‌های فناوری و محققان را تقویت کنید تا استانداردها و راه‌حل‌های مشترکی برای مقابله با دستکاری توسعه دهید.

- **اقدامات شفافیت:**
  - از پلتفرم‌ها بخواهید در مورد ترویج محتوا، هدف‌گیری تبلیغات یا توصیه‌های الگوریتمی شفاف باشند.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت کمپین‌های دستکاری را تشخیص، مهار و کاهش دهید.
- **هوش تهدید پیش‌گیرانه:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی برای نظارت بر روندهای نوظهور دستکاری و تکنیک‌های تطبیقی استفاده کنید.
- **احراز هویت محتوا:**
  - مکانیسم‌های تأیید مبتنی بر بلاک‌چین یا رمزنگاری را پیاده‌سازی کنید تا اصالت محتوای رسانه‌ای تضمین شود.

۲-۵۳ جَمینگ

حمله جَمینگ[1] شامل مختل‌کردن کانال‌های ارتباطی با ارسال سیگنال‌های تداخلی است که باعث می‌شود انتقال داده‌های قانونی غیرممکن یا غیرقابل اعتماد شود. در حمله جَمینگ مبتنی بر هوش مصنوعی، از هوش مصنوعی برای افزایش دقت، انعطاف‌پذیری و کارایی تکنیک‌های سنتی جَمینگ استفاده می‌شود. با بهره‌گیری از یادگیری ماشین و یادگیری عمیق، مهاجمان می‌توانند ترافیک شبکه را تحلیل کنند، فرکانس‌های بهینه برای جَمینگ را پیش‌بینی کنند و استراتژی‌های خود را به‌طور پویا تنظیم کنند تا از تشخیص یا اقدامات متقابل فرار کنند.

این نوع حمله به‌خصوص در شبکه‌های بی‌سیم، اکوسیستم‌های اینترنت اشیا (IoT) و سیستم‌های زیرساخت حیاتی خطرناک است، جایی که ارتباطات قابل اعتماد ضروری هستند. حملات جَمینگ مبتنی بر هوش مصنوعی می‌توانند از دفاع‌های سنتی مانند پرش فرکانسی یا فیلتر سیگنال عبور کنند و با هدف‌گیری هوشمندانه آسیب‌پذیری‌ها در زمان واقعی، اختلال ایجاد کنند.

۲-۵۳-۱ ویژگی‌های کلیدی

- **خودکارسازی:**
  - هوش مصنوعی فرآیند شناسایی کانال‌های ارتباطی، تولید سیگنال‌های تداخلی و اجرای

1. Jamming

حملات را خودکار می‌کند و تلاش دستی را کاهش می‌دهد.

- **انعطاف‌پذیری:**
  - حملات مبتنی بر هوش مصنوعی به‌طور پویا تکامل می‌یابند و با تغییرات در پیکربندی شبکه، اقدامات دفاعی یا شرایط محیطی سازگار می‌شوند.
- **پنهان‌کاری:**
  - با تقلید از سیگنال‌های قانونی یا رمزنگاری payloadها تا زمان اجرا، هوش مصنوعی اطمینان می‌دهد که فعالیت‌های جَمینگ کشف نشوند.
- **هدف‌گیری دقیق:**
  - الگوریتم‌های پیشرفته به مهاجمان امکان می‌دهند کانال‌های ارتباطی، دستگاه‌ها یا سیستم‌های خاص را برای حداکثر تأثیر شناسایی کنند.
- **مقیاس‌پذیری:**
  - هوش مصنوعی به مهاجمان امکان می‌دهد چندین شبکه یا سیستم را به‌طور همزمان مختل کنند و آسیب‌های بالقوه را افزایش دهند.

۲-۵۳-۲ تکنیک‌های هوش مصنوعی مورد استفاده

- **یادگیری ماشین:**
  - یادگیری نظارت‌شده: مدل‌ها را روی مجموعه‌داده‌های حملات جَمینگ موفق و ناموفق آموزش می‌دهد تا استراتژی‌ها را بهبود بخشد.
  - یادگیری بدون نظارت: الگوها یا ناهنجاری‌ها در ترافیک بی‌سیم را شناسایی می‌کند تا آسیب‌پذیری‌ها یا فرکانس‌های بهینه برای جَمینگ را کشف کند.
  - یادگیری تقویتی: تاکتیک‌های جَمینگ را با پاداش‌دهی به نتایج موفق و جریمه‌کردن شکست‌ها بهینه می‌کند و امکان بهبود مداوم را فراهم می‌کند.
- **یادگیری عمیق:**
  - شبکه‌های عصبی، به‌ویژه شبکه‌های عصبی بازگشتی و شبکه‌های عصبی پیچشی، وابستگی‌های زمانی پیچیده در سیگنال‌های بی‌سیم را تحلیل می‌کنند تا زمان‌بندی و شدت تداخل را بهبود بخشند.
  - شبکه‌های مولد تخاصمی: سیگنال‌های تداخلی مصنوعی اما واقع‌گرایانه ایجاد می‌کنند تا قابلیت‌های فرار را آزمایش و بهبود بخشند.

- **پردازش سیگنال:**
  - از تبدیل‌های فوریه، تحلیل موجک یا تحلیل مؤلفه‌های اصلی (PCA) برای پیش‌پردازش و استخراج ویژگی‌های معنادار از سیگنال‌های بی‌سیم خام استفاده می‌کند.
- **مدل‌سازی رفتاری:**
  - از روش‌های آماری و یادگیری ماشین برای مدل‌سازی رفتار عادی ارتباطات بی‌سیم استفاده می‌کند تا انحرافات را تشخیص دهد یا فعالیت قانونی را جعل کند.
- **سنجش محیطی:**
  - عوامل محیطی مانند قدرت سیگنال، سطح نویز و مکان دستگاه‌ها را تحلیل می‌کند تا اثربخشی جَمینگ را بهینه کند.

۳-۵۳-۲ دارایی‌های هدف

- **شبکه‌های بی‌سیم:**
  - شبکه‌های Wi-Fi، بلوتوث یا سلولی که در آن‌ها اختلال می‌تواند منجر به قطع سرویس یا از دست‌دادن داده شود.
- **اکوسیستم‌های اینترنت اشیا (IoT):**
  - لوازم خانگی هوشمند، دستگاه‌های پوشیدنی یا سنسورهای صنعتی که برای عملکرد به ارتباطات بی‌سیم وابسته هستند.
- **زیرساخت‌های حیاتی:**
  - شبکه‌های برق، تأسیسات بهداشتی، سیستم‌های حمل‌ونقل یا خدمات اضطراری که در آن‌ها اختلال در ارتباطات می‌تواند آسیب‌های قابل توجهی ایجاد کند.
- **سیستم‌های نظامی و دفاعی:**
  - شبکه‌های ارتباطی، پهپادها یا سیستم‌های ماهواره‌ای که در برابر حملات جَمینگ هدف‌گیری آسیب‌پذیر هستند.
- **ارائه‌دهندگان ارتباطات:**
  - شبکه‌های موبایل، ارائه‌دهندگان خدمات اینترنت (ISP) یا ارائه‌دهندگان VoIP که زیرساخت آن‌ها حجم عظیمی از داده‌های کاربر را مدیریت می‌کنند.

۲-۵۳-۴ آسیب‌پذیری‌ها

- **احراز هویت ضعیف سیگنال:**
  - عدم وجود مکانیسم‌های قوی برای تأیید اصالت سیگنال‌های بی‌سیم، شبکه‌ها را در معرض تداخل قرار می‌دهد.
- **اقدامات ضد جَمینگ ناکافی:**
  - عدم وجود فناوری‌های پیشرفته ضد جَمینگ مانند پرش فرکانسی یا تکنیک‌های طیف گسترده.
- **خطای انسانی:**
  - پیکربندی نادرست فایروال‌ها، روترها یا آنتن‌ها به‌طور ناخواسته آسیب‌پذیری‌هایی را ایجاد می‌کند که توسط جَمینگ مبتنی بر هوش مصنوعی قابل بهره‌برداری هستند.
- **دفاع‌های قدیمی:**
  - استفاده از سیستم‌های قدیمی یا نرم‌افزارهای منسوخ، سازمان‌ها را در برابر تکنیک‌های جَمینگ انطباق‌پذیر آسیب‌پذیر می‌کند.
- **اتکای بیش از حد به خودکارسازی:**
  - سیستم‌هایی که فقط به ابزارهای خودکار برای اعتبارسنجی یا طبقه‌بندی متکی هستند، ممکن است نتوانند سیگنال‌های جَمینگ ظریف یا در حال تکامل را تشخیص دهند.

۲-۵۳-۵ سناریوی تهدید

- **شناسایی:**
  - مهاجم از ابزارهای مبتنی بر هوش مصنوعی برای جمع‌آوری اطلاعات درباره شبکه هدف، از جمله معماری، پروتکل‌های ارتباطی و الگوهای ترافیک معمول استفاده می‌کند.
- **تحلیل و آموزش:**
  - مهاجم مدل‌های یادگیری ماشین را برای تحلیل داده‌های تاریخی ترافیک بی‌سیم و آموزش روی الگوهای رفتار قانونی به‌کار می‌گیرد تا استراتژی‌های جَمینگ را بهبود بخشد.
- **تولید سیگنال جَمینگ:**
  - مهاجم از یادگیری عمیق یا شبکه‌های مولد تخاصمی برای ایجاد سیگنال‌های تداخلی سفارشی‌سازی‌شده متناسب با آسیب‌پذیری‌های شناسایی‌شده استفاده می‌کند.

- **استقرار:**
  - مهاجم استراتژی جَمینگ برنامه‌ریزی‌شده را با ارسال سیگنال‌های تداخلی در فرکانس‌ها، زمان‌بندی‌ها و شدت‌های بهینه اجرا می‌کند.
- **پایداری:**
  - مهاجم سیگنال‌های جَمینگ را بر اساس بازخورد بلادرنگ به‌طور مداوم به‌روزرسانی و سازگار می‌کند تا اختلال بلندمدت را حفظ کند.
- **پاک‌کردن ردپا:**
  - مهاجم لاگ‌ها را تغییر می‌دهد، ردپاها را پاک می‌کند یا از تکنیک‌های دیگر برای جلوگیری از تشخیص و حفظ پنهان‌کاری استفاده می‌کند.

۲-۵۳-۶ نشانه‌های احتمال نفوذ

- **افزایش سطح نویز:**
  - ناهنجاری‌ها در نسبت سیگنال به نویز که از هنجارهای مورد انتظار منحرف می‌شوند.
- **اختلال در ارتباطات:**
  - قطع‌های غیرمنتظره در اتصال، افزایش از دست‌دادن بسته‌ها یا کاهش عملکرد در شبکه‌های بی‌سیم.
- **شاخص‌های رفتاری:**
  - اقدامات ناسازگار با رفتار عادی شبکه، مانند اتصال‌های مکرر یا نرخ خطای غیرعادی.
- **ورودی‌های لاگ:**
  - ورودی‌های مشکوک در لاگ‌های سیستم مربوط به رویدادهای غیرمنتظره، احراز هویت ناموفق یا تلاش‌های دسترسی غیرمجاز.
- **تغییرات غیرمنتظره فرکانس:**
  - انحراف در فرکانس‌های عملیاتی یا تغییرات ناگهانی در پروتکل‌های ارتباطی که ممکن است نشان‌دهنده تلاش‌های جَمینگ باشد.

۲-۵۳-۷ تأثیرات

- **اختلال در عملیات:**
  - وقفه در خدمات یا سیستم‌های حیاتی کسب‌وکار به دلیل کانال‌های ارتباطی مخدوش.

- **ضررهای مالی:**
  - downtime سرویس، کاهش بهره‌وری یا آسیب به اعتبار که منجر به ضرر مالی می‌شود.
- **خطرات ایمنی:**
  - در زیرساخت‌های حیاتی، حملات جَمینگ می‌توانند جان افراد را به خطر بیندازند یا باعث شکست‌های فاجعه‌بار شوند.
- **آسیب به اعتبار:**
  - سازمان‌هایی که از حملات جَمینگ مبتنی بر هوش مصنوعی رنج می‌برند ممکن است اعتماد مشتریان و ارزش برند خود را از دست بدهند.
- **فساد سیستم:**
  - قرار گرفتن طولانی‌مدت در معرض سیگنال‌های تداخلی می‌تواند فریم‌ور یا اجزای سخت‌افزاری دستگاه‌های بی‌سیم را فاسد کند.

۸-۵۳-۲ راه‌های مقابله

- **تشخیص مبتنی بر هوش مصنوعی:**
  - سیستم‌های مبتنی بر هوش مصنوعی را مستقر کنید که قادر به شناسایی آثار ظریف یا ناسازگاری‌ها در سیگنال‌های بی‌سیم هستند که نشان‌دهنده جَمینگ است.
- **پرش فرکانسی:**
  - تکنیک‌های پرش فرکانسی پویا یا طیف گسترده را پیاده‌سازی کنید تا اثربخشی حملات جَمینگ را کاهش دهید.
- **احراز هویت سیگنال:**
  - از روش‌های رمزنگاری برای احراز هویت سیگنال‌های بی‌سیم و تضمین یکپارچگی آن‌ها استفاده کنید.
- **به‌روزرسانی‌های منظم:**
  - تمام نرم‌افزارها، فریم‌ورها و پیکربندی‌ها را به‌روز نگه دارید تا آسیب‌پذیری‌های شناخته‌شده مورد سوءاستفاده در جَمینگ اصلاح شوند.
- **تقسیم‌بندی شبکه:**
  - شبکه‌های بی‌سیم را به بخش‌های کوچک‌تر تقسیم کنید تا گسترش تداخل محدود شده و تأثیر حملات موفق کاهش یابد.

- **آموزش و آگاهی:**
  - کارکنان و کاربران را در مورد شناسایی علائم جَمینگ آموزش دهید و فرهنگ هوشیاری را تقویت کنید.
- **هوش تهدید پیشرفته:**
  - از پلتفرم‌های هوش تهدید مبتنی بر هوش مصنوعی استفاده کنید تا از تکنیک‌های نوظهور جَمینگ و تهدیدات انطباق‌پذیر مطلع شوید.
- **برنامه‌ریزی پاسخ به حوادث:**
  - برنامه‌های پاسخ به حوادث قوی را توسعه و پیاده‌سازی کنید تا به‌سرعت حملات جَمینگ را تشخیص داده، مهار کرده و اثرات آن‌ها را کاهش دهید.
- **دفاع پیشگیرانه:**
  - از یادگیری ماشین تخاصمی برای شبیه‌سازی و آزمایش اقدامات دفاعی در برابر استراتژی‌های در حال تکامل جَمینگ استفاده کنید.
- **اقدامات امنیتی فیزیکی:**
  - تجهیزات ارتباطی حیاتی را با موانع فیزیکی، محافظها یا محفظه‌های مقاوم در برابر دستکاری محافظت کنید تا دسترسی به دستگاه‌های جَمینگ محدود شود.

۳ شناسنامه حملات هوش مصنوعی تخاصمی

۳-۱ مقدمه

سامانه‌های هوش مصنوعی در حال گسترش شتاب‌دار چندساله‌ای در سطح جهان هستند. این سامانه‌ها توسط کشورهای متعددی توسعه داده شده و به طور گسترده در اقتصادهای آن‌ها مستقر شده‌اند، که منجر به ظهور خدمات مبتنی بر هوش مصنوعی برای استفاده مردم در بسیاری از حوزه‌های زندگی، هم واقعی و هم مجازی شده است. هوش مصنوعی به دو دسته گسترده تقسیم می‌شود که بر اساس قابلیت‌هایشان تعریف می‌شوند: هوش مصنوعی پیش‌بین (PredAI)[1] و هوش مصنوعی مولد (GenAI)[2]. با نفوذ این سامانه‌ها در اقتصاد دیجیتال و تبدیل شدن به بخش‌های جدایی‌ناپذیر و ضروری از زندگی روزمره، نیاز به عملکرد امن، قوی و انعطاف‌پذیر آن‌ها افزایش یافته است. این ویژگی‌های عملیاتی، عناصر کلیدی از هوش مصنوعی قابل اعتماد در چارچوب مدیریت ریسک هوش مصنوعی NIST و در طبقه‌بندی قابلیت اعتماد هوش مصنوعی محسوب می‌شوند.

با این حال، علیرغم پیشرفت‌های قابل توجهی که هوش مصنوعی و یادگیری ماشین در حوزه‌های مختلف کاربردی داشته‌اند، این فناوری‌ها نیز در برابر حملات آسیب‌پذیر هستند که می‌توانند منجر به شکست‌های باورنکردنی و عواقب جدی شوند.

به عنوان مثال، در کاربردهای بینایی ماشین هوش مصنوعی پیش‌بین (PredAI) برای تشخیص و دسته‌بندی اشیا، موارد مشهوری از داده‌های متخاصم‌شده در تصاویر ورودی باعث شده‌اند که خودروهای خودران به خط مخالف جاده منحرف شوند. اشتباه در دسته‌بندی تابلوهای «توقف» به عنوان تابلوهای «محدودیت سرعت»، باعث شده است که اشیای مهم از تصاویر ناپدید شوند و حتی افرادی که در محیط‌های امنیتی بالا عینک می‌زنند به اشتباه شناسایی شوند. به طور مشابه، در حوزه پزشکی که تعداد مدل‌های یادگیری ماشین برای کمک به پزشکان بیشتر می‌شود، احتمال نشت رکوردهای پزشکی از مدل‌های یادگیری ماشین وجود دارد که می‌تواند اطلاعات شخصی عمیق را فاش کند.

در هوش مصنوعی مولد (GenAI)، مدل‌های زبانی بزرگ (LLMs)[3] نیز به بخشی جدایی‌ناپذیر از زیرساخت اینترنت و برنامه‌های نرم‌افزاری تبدیل شده‌اند. LLMها برای ایجاد جستجوهای آنلاین قوی‌تر، کمک به توسعه‌دهندگان نرم‌افزار در نوشتن کد و حتی توانمندسازی چت‌بات‌هایی که به خدمات مشتریان کمک می‌کنند، استفاده می‌شوند. LLMها با پایگاه‌های داده و اسناد شرکتی ادغام

1. Predictive AI
2. Generative AI
3. Large Languages Models

شده‌اند تا سناریوهای قدرتمند تولید تقویت‌شده با بازیابی (RAG)[1] را فعال کنند، زمانی که LLMها برای حوزه‌ها و موارد استفاده خاص تنظیم می‌شوند. این سناریوها در واقع سطح حمله جدیدی را برای داده‌های محرمانه و انحصاری شرکت‌ها فراهم می‌کنند.

به استثنای BLOOM و LLaMA ، بیشتر شرکت‌هایی که این مدل‌ها را توسعه می‌دهند، اطلاعات دقیقی درباره مجموعه داده‌هایی که برای ساخت مدل‌های زبانی خود استفاده کرده‌اند منتشر نمی‌کنند، اما این مجموعه داده‌ها به طور غیرقابل اجتناب شامل برخی از اطلاعات شخصی حساس می‌شوند، مانند آدرس‌ها، شماره‌های تلفن و آدرس‌های ایمیل. این موضوع خطرات جدی برای حریم خصوصی کاربران آنلاین ایجاد می‌کند. هرچه قطعه‌ای از اطلاعات بیشتر در یک مجموعه داده ظاهر شود، احتمال نشت آن توسط مدل در پاسخ به پرس‌وجوهای تصادفی یا طراحی‌شده خاص بیشتر می‌شود. این موضوع می‌تواند ارتباطات اشتباه و مضر را تداوم بخشد که عواقب مخربی برای افراد درگیر دارد و نگرانی‌های اضافی در زمینه امنیت و ایمنی به همراه دارد. مهاجمان همچنین می‌توانند داده‌های آموزشی سامانه‌های هوش مصنوعی پیش‌بین (PredAI) و تولیدی (GenAI) را دستکاری کنند، که این موضوع باعث می‌شود سامانه هوش مصنوعی آموزش‌دیده بر روی این داده‌ها در برابر حملات آسیب‌پذیر شود. جمع‌آوری داده‌های آموزشی از اینترنت همچنین امکان آلوده‌سازی داده‌ها در مقیاس گسترده را به هکرها می‌دهد تا آسیب‌پذیری‌هایی ایجاد کنند که منجر به نقض امنیت در مراحل بعدی شوند.

با رشد اندازه مدل‌های یادگیری ماشین، بسیاری از سازمان‌ها به مدل‌های آموزش‌دیده از قبل متکی شده‌اند که می‌توانند مستقیماً استفاده شوند یا با مجموعه داده‌های جدید تنظیم دقیق شوند تا وظایف مختلفی را فعال کنند. این موضوع فرصت‌هایی برای تغییرات خرابکارانه در مدل‌های آموزش‌دیده از قبل ایجاد می‌کند، مانند قرار دادن تروجان‌ها برای اینکه مهاجمان بتوانند دسترسی به مدل را مختل کنند، پردازش نادرست اجباری کنند یا داده‌ها را در صورت دستور نشت دهند.

از نظر تاریخی، فناوری هوش مصنوعی خاص حالت‌های ورودی (مانند متن، تصویر، گفتار، داده‌های جدولی) در سامانه‌های PredAI و GenAI ظهور کرده است که هر کدام مستعد حملات خاص حوزه خود هستند. به عنوان مثال، رویکردهای حمله علیه وظایف طبقه‌بندی تصویر به طور مستقیم به حملات علیه مدل‌های پردازش زبان طبیعی ترجمه نمی‌شوند. اخیراً، معماری‌های ترانسفورمر که به طور گسترده در پردازش زبان طبیعی استفاده می‌شوند، کاربردهایی در حوزه بینایی ماشین نشان داده‌اند. علاوه بر این، یادگیری چند حالته[2] پیشرفت‌های هیجان‌انگیزی در بسیاری از وظایف، از جمله وظایف تولیدی و طبقه‌بندی، داشته است و تلاش‌هایی برای استفاده از یادگیری چند حالته به

1. Retrieval Augmented Generation
2. Multimodal ML

عنوان یک راهکار بالقوه برای کاهش حملات تک‌حالته انجام شده است. با این حال، حملات قدرتمند همزمان علیه تمام حالات در یک مدل چند حالته نیز ظهور کرده‌اند.

به طور بنیادی، روش‌شناسی یادگیری ماشین مورد استفاده در سامانه‌های هوش مصنوعی مدرن مستعد حملات از طریق APIهای عمومی است که مدل را در معرض قرار می‌دهند، و همچنین در برابر پلتفرم‌هایی که روی آن‌ها مستقر شده‌اند. این گزارش بر روی مورد اول تمرکز دارد و مورد دوم را در حوزه طبقه‌بندی‌های سنتی امنیت سایبری در نظر می‌گیرد. برای حملات علیه مدل‌ها، مهاجمان می‌توانند محرمانگی و حفاظت از حریم خصوصی داده‌ها و مدل را با استفاده ساده از رابطهای عمومی مدل و ارائه ورودی‌های داده‌ای که در محدوده قابل قبول هستند، نقض کنند. از این نظر، چالش‌های پیش‌روی یادگیری ماشین تطبیقی مشابه چالش‌های رمزنگاری مدرن است. رمزنگاری مدرن بر الگوریتم‌هایی استوار است که از لحاظ نظری اطلاعاتی ایمن هستند. بنابراین، مردم فقط باید بر اجرای قوی و ایمن آن‌ها تمرکز کنند—که این کار هم کوچک نیست. برخلاف رمزنگاری، هیچ اثبات امنیت نظری اطلاعاتی برای الگوریتم‌های یادگیری ماشین گسترده‌ای وجود ندارد. علاوه بر این، نتایج غیرممکن‌بودن نظری اطلاعاتی شروع به ظهور در متون علمی کرده‌اند. که محدودیت‌هایی بر مؤثر بودن تکنیک‌های کاهش خطر گسترده‌استفاده شده تعیین می‌کنند. در نتیجه، بسیاری از پیشرفت‌ها در توسعه راهکارهای کاهش خطر علیه انواع مختلف حملات تمایل به ماهیت تجربی و محدود دارند.

مبتنی بر دسته‌بندی NIST در حوزه حملات خصمانه در این پروژه حملات تخاصمی بر مدل‌های هوش مصنوعی پیش‌بین در جدول ۳ نمایش داده شده است.

جدول ۳: حملات تخاصمی بر هوش مصنوعی پیش‌بین

مسمووم سازی	مسمومیت مدل هوش مصنوعی
	مسمومیت با برچسب تمیز
	مسمومیت هدفمند
	مسمومیت درب پشتی
	تزریق مستقیم درخواست
حملات حریم خصوصی	بازسازی داده
	استنتاج عضویت
	استنتاج ویژگی
	استخراج مدل
گریز یا نمونه‌های متخاصم	

در خصوص حملات تخاصمی بر مدل‌های مولد علاوه بر حملات که در هوش مصنوعی پیش‌بین

وجود داشته که در مدل‌های مولد نیز قابلیت اجرایی شدند دارند، سه حمله تزریق مستقیم درخواست، تزریق غیرمستقیم درخواست و حمله مبتنی بر RAG نیز در این پروژه مورد بررسی قرار گرفته است.

همانند حملات شناسایی شده در هوش مصنوعی تهاجمی، برای تمامی حملات تخاصمی یک شناسنامه مبتنی بر ۷ مؤلفه تهیه شده است. برای هر یک از حملات ابتدا توضیح مختصری بیان شده و سپس موارد زیر مشخص و شرح داده شده است:

- **ویژگی‌های کلیدی:**
  - این بخش به معرفی ویژگی‌هایی هر یک از تهدیدات و حملات می‌پردازد و وجوه تمایز آن را با سایر حملات و تهدیدات باین می‌کند.
- **دارایی‌های هدف:**
  - این بخش اهداف حملات در یک سازمان یا یک سامانه مبتنی بر هوش مصنوعی را بیان می‌کند.
- **آسیب‌پذیری‌ها:**
  - بیانگر نقاط ضعف داخلی سامانه‌های هوش مصنوعی یا محیط عملیاتی و استقرار آن‌هاست که مهاجمان با سوءاستفاده از آن‌ها اقدام به اجرای حملات خود می‌کنند.
- **سناریو حمله:**
  - در این بخش سناریو و مراحل اجرای اصلی حملات بیان می‌گردد. سناریو و مراحل اجرای بیان شده، پایه‌ای است که با استفاده از آن می‌توان سنایو و مراحل اجرای حمله در بخش کاربردی یا بر روی سامانه خاصی که از مدل‌های هوش مصنوعی استفاده می‌کند را بیان کرد.
- **نشانه‌های احتمال نفوذ:**
  - در این بخش فهرستی از مواردی که در صورت بروز احتمال به وقوع پیوستن و یا در حال وقوع بودن حمله را مشخص می‌کند بیان می‌کند. بدیهی است که ممکن است نشانه‌های مشترکی میان حملات گوناگون وجود داشته باشد.
- **تأثیرات:**
  - این بخش به تبعات، پیامدها و تأثیراتی که در صورت روز موفق حملات اتفاق میفتد اختصاص دارد. این پیامدها می‌تواند در ابعاد سامانه یا محصول، سازمان یا در بعد فردی و اجتماعی باشد.
- **راه‌های مقابله:**
  - در این بخش مجموعه ای از راهبردها و اقدامات جهت مقابله و کاهش محاطرات هر یک از

حملات بیان می‌گرد. این بخش نیز ممکن است در برخی از مصادیق میان حملات مختلف مشترک باشد.

در جدول ۴ نگاشت میان حملات فوق و تاکتیک‌های ۱۴ گانه MITRE نمایش داده شده است.

جدول ۴: نگاشت حملات هوش مصنوعی تخاصمی با تاکتیک‌های ۱۴ گانه MITRE

تاکتیک	توضیحات تاکتیک	حملات مرتبط با هوش مصنوعی
جمع‌آوری اطلاعات (Reconnaissance)	جمع‌آوری اطلاعات برای برنامه‌ریزی عملیات آینده.	استخراج مدل بازسازی داده استنتاج عضویت استنتاج ویژگی
توسعه منابع (Resource Development)	ایجاد منابع برای حمایت از عملیات.	مسمومیت مدل هوش مصنوعی مسمومیت با برچسب تمیز مسمومیت هدفمند مسمومیت درب پشتی مسمومیت در دسترس‌پذیری گریز یا نمونه‌های متخاصم تزریق غیرمستقیم درخواست حمله مبتنی بر RAG
دسترسی اولیه (Initial Access)	کسب اولین پایگاه در سامانه یا شبکه هدف.	مسمومیت مدل هوش مصنوعی مسمومیت هدفمند مسمومیت درب پشتی مسمومیت در دسترس‌پذیری تزریق مستقیم درخواست تزریق غیرمستقیم درخواست حمله مبتنی بر RAG
اجرا (Execution)	اجرای کد مخرب در سامانه هدف.	گریز یا نمونه‌های متخاصم تزریق مستقیم درخواست
پایداری (Persistence)	حفظ پایگاه در محیط هدف.	مسمومیت مدل هوش مصنوعی مسمومیت با برچسب تمیز مسمومیت درب پشتی
ارتقاء دسترسی (Privilege Escalation)	کسب مجوزهای سطح بالاتر در سامانه.	-
دور زدن دفاع (Defense Evasion)	اجتناب از تشخیص توسط مکانیزم‌های امنیتی.	مسمومیت با برچسب تمیز مسمومیت هدفمند مسمومیت در دسترس‌پذیری گریز یا نمونه‌های متخاصم تزریق مستقیم درخواست تزریق غیرمستقیم درخواست حمله مبتنی بر RAG

تاکتیک	توضیحات تاکتیک	حملات مرتبط با هوش مصنوعی
دسترسی به اعتبارنامه (Credential Access)	سرقت اعتبارنامه‌ها برای دسترسی به سامانه‌ها.	-
کشف (Discovery)	کشف محیط هدف برای جمع‌آوری اطلاعات.	-
حرکت جانبی (Lateral Movement)	حرکت در محیط هدف برای دسترسی به سامانه‌های اضافی.	-
جمع‌آوری داده‌ها (Collection)	جمع‌آوری داده‌های مورد علاقه از هدف.	استخراج مدل بازسازی داده استنتاج عضویت استنتاج ویژگی
فرمان و کنترل (Command and Control)	ارتباط با سامانه‌های مختل‌شده برای کنترل آن‌ها.	-
استخراج داده‌ها (Exfiltration)	سرقت داده‌ها از هدف.	استخراج مدل بازسازی داده استنتاج عضویت استنتاج ویژگی
تأثیر (Impact)	دستکاری، قطع یا تخریب سامانه‌ها یا داده‌های هدف.	مسمومیت مدل هوش مصنوعی مسمومیت با برچسب تمیز مسمومیت هدفمند مسمومیت درب پشتی مسمومیت در دسترس‌پذیری استخراج مدل گریز یا نمونه‌های متخاصم بازسازی داده استنتاج عضویت حمله مبتنی بر RAG استنتاج ویژگی تزریق مستقیم درخواست تزریق غیرمستقیم درخواست

بسیاری از حملات تخاصمی، مانند حملات مسموم‌سازی یا استنتاج، شامل چندین تاکتیک هستند زیرا شامل آماده‌سازی (توسعه منابع)، بهره‌برداری (دسترسی اولیه یا اجرا) و نتایج (تأثیر یا خروج داده) می‌شوند. در نگاشت حاضر اهداف مرتبطتر اولویت‌بندی شده‌اند اما همپوشانی‌های احتمالی نیز اعمال شده است.

این حملات سامانه‌ها و راهکارهای هوش مصنوعی را مورد هدف قرار می‌دهند و نه زیرساخت‌های سنتی فناوری اطلاعات، بنابراین تاکتیک‌هایی مانند حرکت جانبی یا فرمان و کنترل ممکن است کمتر مستقیماً قابل اعمال باشند مگر اینکه حمله بخشی از یک کمپین گسترده‌تر باشد و همانطور که گفته شد چون هدف اصلی خود سامانه و راهکار مبتنی بر هوش مصنوعی است بنابراین تمامی حملات

تاکتیک تأثیر را شامل می‌شوند. در ادامه بر مبنای ترتیب مشخص‌شده در جدول ۴ شناسنامه حملات مشخص شده است.

۳-۲ استخراج مدل

استخراج مدل[1]های هوش مصنوعی یک نگرانی رو به رشد در زمینه یادگیری ماشین و هوش مصنوعی است، زیرا خطرات قابل توجهی برای فناوری‌های مالکیتی و داده‌های حساس به همراه دارد. با افزایش استقرار مدل‌های یادگیری ماشین در برنامه‌های مختلف (مالی و بهداشت و درمان تا سامانه‌های خودران) مهاجمان در حال یافتن راه‌هایی برای سوءاستفاده از این مدل‌ها برای مقاصد بدخواهانه هستند.

استخراج مدل معمولاً زمانی رخ می‌دهد که یک مهاجم با یک مدل یادگیری ماشین از طریق رابط برنامه‌نویسی کاربردی (API) تعامل می‌کند. با پرسش‌های نظام‌مند و طراحی‌شده دقیق، مهاجمان می‌توانند اطلاعات کافی را جمع‌آوری کنند تا مدل مشابهی را بازسازی کنند یا به بینش‌هایی در مورد معماری و عملکرد آن دست یابند. این فرایند می‌تواند منجر به تکثیر غیرمجاز مدل شود که ممکن است مزیت رقابتی خالق اصلی را تضعیف کند.

۳-۲-۱ ویژگی‌های کلیدی

در اینجا ویژگی‌های اصلی مرتبط با استخراج مدل آورده شده است:

- **تعامل مبتنی بر پرسش:**
  - مهاجم از طریق پرسش‌ها با مدل تعامل می‌کند و معمولاً از یک API استفاده می‌کند. این روش به مهاجمان اجازه می‌دهد خروجی‌ها را بر اساس مجموعه‌ای از ورودی‌ها جمع‌آوری کرده و برای استنتاج رفتار مدل تجزیه و تحلیل کنند.
- **تحلیل خروجی:**
  - استخراج‌کنندگان خروجی‌های تولید شده توسط مدل در پاسخ به پرسش‌هایشان را تجزیه و تحلیل می‌کنند. با تغییر نظام‌مند ورودی‌ها، مهاجمان می‌توانند الگوها، مرزهای تصمیم‌گیری و ویژگی‌های کلیدی استفاده شده در مدل را شناسایی کنند.
- **بازسازی پارامترهای مدل:**
  - مهاجمان سعی می‌کنند معماری و پارامترهای مدل را مهندسی معکوس کنند. این امر می‌تواند شامل تقریب مدل اصلی با استفاده از داده‌های خروجی جمع‌آوری شده برای ایجاد یک مدل مشابه باشد.

1. Model Extraction

- **وابستگی به داده:**
  - موفقیت استخراج مدل معمولاً به مقدار و تنوع داده‌های موجود از طریق پرسش‌ها بستگی دارد. پرسش‌های گسترده‌تر و متنوع‌تر می‌توانند منجر به تقریبات بهتری از مدل اصلی شوند.
- **ویژگی‌های مدل هدف:**
  - پیچیدگی و نوع مدل هدف می‌تواند بر موفقیت استخراج تأثیر بگذارد. مدل‌های ساده‌تر ممکن است استخراج آن‌ها آسان‌تر باشد، در حالی که مدل‌های پیچیده با معماری‌های قوی‌تر می‌توانند چالش‌های بیشتری ایجاد کنند.
- **یادگیری از بازخورد:**
  - مهاجمان ممکن است راهبردهای پرسش‌گری خود را بر اساس بازخوردی که از مدل دریافت می‌کنند، تنظیم کنند. این رویکرد تطبیقی توانایی آن‌ها را برای استخراج اطلاعات مفید به‌طور مؤثر افزایش می‌دهد.
- **آسیب‌پذیری نسبت به بیش‌برازش:**
  - مدل‌های استخراج شده ممکن است به خروجی‌های خاصی که روی آن‌ها آموزش دیده‌اند، بیش‌برازش کنند که توانایی‌های تعمیم آن‌ها را محدود می‌کند. این موضوع می‌تواند یک شمشیر دو لبه باشد، زیرا همچنین ممکن است منجر به شناسایی آسان‌تر محدودیت‌های مدل استخراج شده شود.
- **پتانسیل نشت اطلاعات:**
  - مدل‌هایی که بر روی داده‌های حساس آموزش دیده‌اند می‌توانند به‌طور ناخواسته اطلاعاتی را از طریق خروجی‌های خود نشت کنند. مهاجمان می‌توانند از این نشت برای استخراج اطلاعات محرمانه موجود در مدل بهره‌برداری کنند.
- **پردازش دسته‌ای:**
  - مهاجمان ممکن است به‌جای درخواست‌های فردی، دسته‌ای از پرسش‌ها ارسال کنند تا حداکثر اطلاعات را کسب کنند. این می‌تواند فرآیند استخراج را تسریع کرده و بینش‌های قوی‌تری از رفتار مدل به دست دهد.
- **استفاده از نمونه‌های متخاصم:**
  - مهاجمان ممکن است از تکنیک‌های متخاصم برای بهبود پرسش‌های خود استفاده کنند. این امر می‌تواند به افزایش اثربخشی استخراج کمک کند و نقاط ضعف مدل را مورد

آزمایش قرار دهد.

۳-۲-۲ دارایی‌های هدف

در زمینه استخراج مدل‌های هوش مصنوعی، چندین دارایی کلیدی به‌طور معمول توسط مهاجمان هدف قرار می‌گیرد که به دنبال تکثیر یا بهره‌برداری از مدل‌های یادگیری ماشین هستند. درک این دارایی‌های هدف می‌تواند به سازمان‌ها کمک کند تا بهتر آماده شوند و از مالکیت معنوی خود محافظت کنند. در اینجا دارایی‌های اصلی در معرض خطر آورده شده است:

- **مدل‌های یادگیری ماشین:**
  - دارایی اصلی که استخراج می‌شود و شامل معماری مدل، پارامترها و نمایندگی‌های یادگرفته‌شده است. دسترسی غیرمجاز به مدل می‌تواند منجر به تکثیر، سوءاستفاده یا بهره‌برداری از الگوریتم‌های مالکیتی شود.
- **داده‌های آموزشی:**
  - داده‌هایی که برای آموزش مدل یادگیری ماشین استفاده می‌شوند و ممکن است شامل اطلاعات حساس یا مالکیتی باشند. مهاجمان می‌توانند اطلاعاتی درباره مجموعه داده‌های آموزشی استنباط کنند که ممکن است منجر به نشت داده یا نقض حریم خصوصی شود.
- **وزن‌ها و پارامترهای مدل:**
  - وزن‌ها و پارامترها مقادیر خاصی هستند که رفتار مدل را تعریف می‌کنند. وزن‌های استخراج شده می‌توانند برای ایجاد یک مدل تکراری که به‌طور مشابه رفتار می‌کند، استفاده شوند و مزیت‌های رقابتی را تضعیف کنند.
- **تکنیک‌های مهندسی ویژگی:**
  - روش‌ها و تکنیک‌های استفاده شده برای انتخاب و تبدیل ویژگی‌های ورودی برای مدل را شامل می‌شوند. اگر مهاجمان بفهمند که ویژگی‌ها چگونه مهندسی شده‌اند، می‌توانند عملکرد مدل را تکرار یا بهبود بخشند.
- **منطق کسب‌وکار و فرآیندهای تصمیم‌گیری:**
  - منظور قوانین و منطق زیرین که مدل برای انجام پیش‌بینی‌ها یا تصمیم‌گیری‌ها استفاده می‌کند، است. آگاهی از فرآیند تصمیم‌گیری مدل می‌تواند منجر به بهره‌برداری یا دستکاری نتایج برای مقاصد بدخواهانه شود.
- **APIها و رابطهای نقطه پایانی:**
  - ضعف در امنیت رابطهایی که از طریق آن‌ها به مدل دسترسی پیدا می‌شود و پرسش‌ها

انجام می‌شود می‌تواند برای تسهیل تلاش‌های استخراج مورد سوءاستفاده قرار گیرد و دسترسی بیشتری به عملکردهای مدل فراهم کند.

- **مستندات و فراداده مدل:**
  - مربوط به معماری مدل، فرآیندهای آموزشی و معیارهای عملکرد و به عبارتی مستندات دقیق که می‌تواند اطلاعاتی را در مورد نحوه کار مدل به مهاجمان ارائه دهد و استخراج را تسهیل کند.
- **داده‌ها و تعاملات کاربر:**
  - داده‌های جمع‌آوری شده از کاربران در حال تعامل با مدل، که ممکن است شامل اطلاعات شخصی باشد. استخراج این داده‌ها می‌تواند منجر به نقض حریم خصوصی و سوءاستفاده احتمالی از اطلاعات کاربر شود.
- **منابع محاسباتی:**
  - درک نیازهای منابع منظور زیرساخت و منابعی که برای آموزش و استقرار مدل استفاده می‌شود می‌تواند به مهاجمان اجازه دهد تا پیاده‌سازی‌های خود را بهینه کنند.
- **معیارهای عملکرد مدل:**
  - اندازه‌گیری‌های که مؤثر بودن مدل، مانند دقت، صحت و نزخ یادآوری را نشان می‌دهند موجب بینش در مورد عملکرد مدل می‌شود و به مهاجمان کمک کند تا مدل‌هایی طراحی کنند که به‌طور مستقیم با مدل اصلی رقابت یا از آن پیشی بگیرند.

۳-۲-۳ آسیب‌پذیری‌ها

در اینجا آسیب‌پذیری‌های کلیدی که می‌توانند در حین استخراج مدل هدف قرار گیرند، آورده شده است:

- **افشای API:**
  - APIهای عمومی می‌توانند توسط هر کسی، از جمله بازیگران بدخواه، مورد پرسش قرار گیرند. عدم کنترل دسترسی مناسب یا محدودیت نرخ می‌تواند منجر به پرسش‌های زیاد شود که این امر باعث می‌شود جمع‌آوری اطلاعات برای مهاجمان آسان‌تر شود.
- **عدم اعتبارسنجی ورودی:**
  - اعتبارسنجی ناکافی از پرسش‌های ورودی می‌تواند به دشمنان اجازه دهد ورودی‌های طراحی‌شده را ارسال کنند. این امر می‌تواند منجر به رفتار غیرمنتظره مدل شود و بینش‌هایی در مورد عملکرد آن را فاش کند.

- **سادگی مدل:**
  - همانظور که گفته شده مدل‌های ساده‌تر نسبت به مدل‌های پیچیده‌تر استخراج آن‌ها آسان‌تر است. اگر یک مدل به اندازه کافی قوی نباشد، مهاجمان می‌توانند به راحتی ساختار و پارامترهای آن را استنباط کنند.
- **خروجی قابل پیش‌بینی:**
  - مدل‌هایی که خروجی‌های ثابت یا به‌راحتی قابل تفسیر ارائه می‌دهند، می‌توانند بیشتر در معرض استخراج باشند. مهاجمان می‌توانند از الگوهای خروجی برای استنتاج فرآیند تصمیم‌گیری مدل بهره‌برداری کنند.
- **عدم تشخیص ناهنجاری کافی:**
  - عدم نظارت بر الگوهای پرسش غیرمعمول می‌تواند به تلاش‌های استخراج اجازه دهد تا بدون شناسایی بمانند. بدون هشدارهای بلادرنگ برای فعالیت‌های غیرعادی، مهاجمان می‌توانند تلاش‌های خود را ادامه دهند بدون اینکه شناسایی شوند.
- **بیش‌برازش:**
  - مدل‌هایی که به داده‌های آموزشی خود بیش‌برازش می‌شوند، ممکن است رفتار پیش‌بینی‌شده‌ای نشان دهند. مهاجمان می‌توانند از این رفتارهای پیش‌بینی‌شده برای ساخت پرسش‌های مؤثر برای استخراج استفاده کنند.
- **نشت داده:**
  - مدل‌هایی که بر روی داده‌های حساس آموزش دیده‌اند ممکن است به‌طور ناخواسته اطلاعاتی را از طریق خروجی‌های خود نشت کنند. مهاجمان می‌توانند از این نشت برای بازسازی مجموعه داده‌های حساس یا درک داده‌های آموزشی مدل بهره‌برداری کنند.
- **اقدامات امنیتی ناکافی:**
  - سازوکارهای ضعیف احراز هویت و مجوز می‌تواند به دسترسی غیرمجاز به مدل اجازه دهد. شیوه‌های امنیتی ضعیف می‌تواند استخراج داده و پرسش‌های غیرمجاز را تسهیل کند.
- **نسخه‌بندی مدل:**
  - عدم مدیریت نسخه‌های مختلف مدل می‌تواند به عدم انسجام در امنیت منجر شود. مدل‌های قدیمی‌تر یا کمتر ایمن ممکن است هنوز قابل دسترسی باشند و هدف آسان‌تری برای استخراج فراهم کنند.

- **محدودیت در تبیین مدل:**
  - مدل‌هایی که تبیین‌پذیری ندارند می‌توانند برای دفاع کردن دشوارتر باشند. اگر کاربران یا مدافعان نتوانند بفهمند که چگونه تصمیمات گرفته می‌شود، شناسایی نقاط ضعف احتمالی دشوار می‌شود.

۳-۲-۴ سناریو تهدید

در یک حمله استخراج مدل، مهاجم با یک مدل یادگیری ماشین که غالباً به‌عنوان یک سرویس ارائه شده است، تعامل می‌کند و با ارسال مجموعه‌ای از پرسش‌های طراحی‌شده به‌دقت، پاسخ‌های مدل به این پرسش‌ها اطلاعات کافی را برای تقریب‌زدن مرزهای تصمیم‌گیری مدل یا بازسازی عملکرد آن فراهم می‌کند. این می‌تواند به تکثیر غیرمجاز مدل‌های اختصاصی یا افشای داده‌های آموزشی حساس که به‌طور غیرمستقیم توسط مدل آموخته شده‌اند، منجر شود.

- **شناسایی:**
  - مهاجم اطلاعاتی درباره مدل جمع‌آوری می‌کند، مانند نوع استفاده (طبقه‌بندی، رگرسیون)، ویژگی‌های ورودی و در دسترس بودن API. این ممکن است شامل درک مستندات عمومی یا نظارت بر الگوهای استفاده معمول باشد.
- **تولید پرسش:**
  - مهاجم مجموعه‌ای از ورودی‌ها را برای پرسش از مدل تولید می‌کند. این ورودی‌ها به‌گونه‌ای طراحی شده‌اند که حداکثر اطلاعات را درباره مرزهای تصمیم‌گیری مدل بیاموزند. تکنیک‌ها می‌توانند از پرسش‌های تصادفی تا ورودی‌های به‌دقت انتخاب‌شده که حداکثر اطلاعات را درباره رفتار مدل به‌دست می‌آورند، متغیر باشند.
- **جمع‌آوری پاسخ‌ها:**
  - مهاجم این پرسش‌ها را به مدل ارسال کرده و خروجی‌ها یا پیش‌بینی‌های ارائه‌شده توسط مدل را جمع‌آوری می‌کند. این پاسخ‌ها به مهاجم کمک می‌کند تا منطق داخلی مدل یا پارامترهای آموخته‌شده آن را استنباط کند.
- **تحلیل و بازسازی مدل:**
  - با استفاده از جفت‌های ورودی-خروجی جمع‌آوری‌شده، مهاجم از الگوریتم‌های یادگیری ماشین برای تقریب‌زدن عملکرد مدل هدف استفاده می‌کند. این ممکن است شامل تکنیک‌هایی باشد که یک مدل جدید را بر روی مجموعه داده‌های جمع‌آوری‌شده آموزش می‌دهد تا رفتار مدل هدف را تقلید کند.

- **ارزیابی و تصحیح:**
  - مهاجم مدل بازسازی‌شده را با تکرار پرسش‌های اضافی و تنظیم دقیق برای بهبود دقت تصحیح می‌کند و تلاش می‌کند تا آن را به‌گونه‌ای مشابه با مدل اصلی کند.

۳-۲-۵ شانه‌های احتمال نفوذ

در این بخش برخی از نشانه‌های کلیدی آورده شده است که ممکن است نشان‌دهنده تلاش یا وقوع استخراج مدل باشند:

- **الگوهای غیرمعمول پرسش API:**
  - ناهنجاری‌ها در فرکانس یا ساختار پرسش‌های API، مانند افزایش ناگهانی درخواست‌ها یا الگوهای تکراری بسیار زیاد و افزایش قابل توجه در تعداد پرسش‌ها از یک آدرس IP واحد یا یک محدوده کوچک از آدرس‌ها در یک بازه زمانی کوتاه را می‌توان از جمله نشانه‌های نفوذ دانست.
- **حجم بالای پرسش‌های ورودی مشابه:**
  - مجموعه‌ای از پرسش‌هایی که به‌هم نزدیک یا تنها کمی متفاوت هستند و ممکن است نشان‌دهنده تلاش برای مهندسی معکوس مدل باشد. لاگ‌هایی که نشان‌دهنده تعداد زیادی پرسش مشابه هستند، به‌ویژه اگر به موارد خاص یا نقاط ضعف شناخته‌شده هدف‌گیری شده باشند.
- **دسترسی از آدرس‌های IP ناشناخته:**
  - پرسش‌هایی که از آدرس‌های IP ناآشنا یا مشکوک منشاء می‌گیرند، به‌ویژه از مناطقی که سازمان در آنجا فعالیت نمی‌کند. تحلیل جغرافیایی که نشان‌دهنده دسترسی از مکان‌های غیرمعمول جغرافیایی یا رنج‌های IP شناخته‌شده بدخواه است.
- **افزایش تاخیر یا کاهش عملکرد:**
  - کاهش قابل توجه در زمان پاسخ API یا عملکرد کلی سامانه به دلیل پرسش‌های زیاد . ابزارهای نظارت بر سامانه که گزارش کاهش در معیارهای عملکرد را می‌دهند و ممکن است با افزایش حجم پرسش‌ها همبستگی داشته باشد.
- **رفتار غیرعادی کاربر:**
  - الگوهای غیرمعمول رفتار از کاربران مشروع، مانند دسترسی به مدل از چندین دستگاه یا مکان در یک بازه زمانی کوتاه. لاگ‌های فعالیت کاربر که تغییرات ناگهانی در الگوهای دسترسی یا تلاش‌های ورود به سامانه نشان می‌دهند که از الگوهای مرسوم انحراف دارند.

- **پرسش مکرر به نقاط پایانی حساس:**
  - افزایش پرسش‌ها به نقاط پایانی خاص مدل که خروجی‌های حساس یا کارکردهای بحرانی را ارائه می‌دهند. لاگ‌ها که نشان‌دهنده این است که نقاط پایانی حساس به‌طور نامتناسبی نسبت به سایر نقاط پایانی مورد دسترسی قرار می‌گیرند.
- **تلاش‌های ناموفق برای احراز هویت:**
  - چندین تلاش ناموفق برای ورود که ممکن است نشان‌دهنده تلاش یک مهاجم برای دسترسی غیرمجاز باشد. لاگ‌های امنیتی که نشان‌دهنده تلاش‌های مکرر ناموفق برای احراز هویت از همان یا آدرس‌های IP مختلف هستند.
- **تغییرات غیرمنتظره در رفتار مدل:**
  - تغییر ناگهانی در الگوهای خروجی مدل یا فرآیندهای تصمیم‌گیری، که ممکن است نشان‌دهنده این باشد که مدل به خطر افتاده یا بازسازی شده است. نظارت بر معیارهای عملکرد مدل که پیش‌بینی‌ها یا طبقه‌بندی‌های غیرمنتظره را نشان می‌دهد.
- **لاگ‌های دسترسی به متادیتا:**
  - دسترسی غیرمعمول به مستندات یا متادیتای مرتبط با مدل که ممکن است نشان‌دهنده تلاش‌های شناسایی یک مهاجم باشد. لاگ‌هایی که نشان‌دهنده دسترسی به مستندات یا فایل‌های پیکربندی هستند که معمولاً توسط کاربران مورد دسترسی قرار نمی‌گیرند.
- **هشدارها از ابزارهای نظارت امنیتی:**
  - هشدارهای خودکار تولید شده توسط سامانه‌های امنیتی که فعالیت یا الگوهای غیرمعمول مرتبط با استخراج مدل را تشخیص می‌دهند. اعلان‌ها از سامانه‌های تشخیص نفوذ (IDS) یا فایروال‌های برنامه وب که نشان‌دهنده تلاش‌های احتمالی استخراج است.

۳-۲-۶ تأثیرات

استخراج مدل‌های هوش مصنوعی می‌تواند تأثیرات قابل توجه و متنوعی بر سازمان‌ها، افراد و چشم‌انداز فناوری داشته باشد. در اینجا برخی از تأثیرات اصلی مرتبط با استخراج مدل‌های هوش مصنوعی آورده شده است:

- **سرقت مالکیت معنوی:**
  - از دست رفتن الگوریتم‌ها و تکنیک‌های اختصاصی که مزیت رقابتی سازمان‌ها را تضعیف می‌کند. رقبا ممکن است مدل دزدیده‌شده را تکرار یا بهبود دهند، که این امر سهم بازار و سودآوری را کاهش می‌دهد.

- **زیان مالی:**
  - زیان‌های مالی مستقیم ناشی از سرقت، دعاوی احتمالی و هزینه‌های مرتبط با کاهش نقص‌ها. سازمان‌ها ممکن است با کاهش درآمد، افزایش هزینه‌های عملیاتی و مخارج مربوط به اقدامات قانونی مواجه شوند.
- **نقض حریم خصوصی داده‌ها:**
  - نشت داده‌های حساس یا اختصاصی که در آموزش مدل‌ها استفاده شده‌اند، می‌تواند منجر به نقض حریم خصوصی شود.سازمان‌ها ممکن است با جریمه‌های نظارتی، آسیب به شهرت و از دست دادن اعتماد مشتری مواجه شوند.
- **آسیب به شهرت:**
  - افشای عمومی حوادث استخراج مدل می‌تواند به اعتبار و شهرت سازمان آسیب برساند. از دست دادن اعتماد مشتری و احتمال اینکه مشتریان بالقوه، رقبا را به سازمان آسیب‌دیده ترجیح دهند.
- **کاهش مزیت رقابتی:**
  - زمانی که رقبا به مدل‌های پیشرفته هوش مصنوعی دسترسی پیدا می‌کنند، موقعیت منحصر به فرد سازمان اصلی تضعیف می‌شود. افزایش رقابت و فشار بر راهبردهای قیمت‌گذاری می‌تواند بر قابلیت بقا در بلندمدت تأثیر بگذارد.
- **افزایش ریسک‌های امنیت سایبری:**
  - سازمان‌ها ممکن است با تهدیدات بیشتری از سوی سایبرمجرمان مواجه شوند که می‌توانند از مدل‌های استخراج‌شده برای مقاصد مخرب استفاده کنند. احتمال حملات بیشتر، از جمله دستکاری داده‌ها یا اجرای حملات متخاصم علیه مدل اصلی.
- **تأثیر بر نوآوری:**
  - استخراج گسترده مدل‌ها ممکن است سرمایه‌گذاری در تحقیق و توسعه هوش مصنوعی را دلسرد کند. سازمان‌ها ممکن است از نوآوری یا اشتراک‌گذاری پیشرفت‌ها خودداری کنند که منجر به کاهش پیشرفت کلی فناوری هوش مصنوعی می‌شود.
- **مسائل قانونی و رعایت مقررات:**
  - سازمان‌ها ممکن است با اقدامات قانونی از سوی طرف‌های آسیب‌دیده یا نهادهای نظارتی به دلیل نقض قوانین حفاظتی یا اطلاعاتی مواجه شوند. افزایش نظارت، دعوی‌های قانونی و نیاز به اقدامات انطباقی بیشتر می‌تواند منابع و توجه را منحرف کند.

- **عواقب منفی برای کاربران:**
  - کاربران نهایی ممکن است در صورتی که داده‌هایشان به خطر بیفتد یا مدل‌ها به‌طور مخرب استفاده شوند، با عواقب منفی مواجه شوند. سوءاستفاده از داده‌های کاربر، می‌تواند منجر به حملات هدفمند یا دستکاری خدمات بر اساس مدل‌های استخراج‌شده شود.
- **تغییر در استانداردهای صنعتی:**
  - حوادث مکرر استخراج مدل ممکن است تغییرات صنعتی در پروتکل‌های امنیتی و استانداردها را تحریک کند. سازمان‌ها ممکن است نیاز به پذیرش تدابیر امنیتی سخت‌گیرانه‌تری داشته باشند که بر نحوه توسعه و استقرار سامانه‌های هوش مصنوعی تأثیر می‌گذارد.

۳-۲-۷ راه‌های مقابله

برای محافظت در برابر استخراج مدل‌های هوش مصنوعی، سازمان‌ها می‌توانند راهبردهای متنوعی را پیاده‌سازی کنند. این راهبردها بر تقویت تدابیر امنیتی، بهبود کنترل‌های دسترسی و ترویج فرهنگ آگاهی تمرکز دارند. در اینجا اقدامات کلیدی برای در نظر گرفتن آورده شده است:

- **تقویت امنیت API:**
  - تعداد درخواست‌هایی که یک کاربر می‌تواند در یک بازه زمانی مشخص ارسال کند محدود شود تا از پرسش‌های بیش از حد جلوگیری شود. هم‌چنین از سازوکارهای احراز هویت قوی، مانند OAuth یا کلیدهای API استفاده شده و اطمینان حاصل شود که کاربران دسترسی مناسبی دارند.
- **اعتبارسنجی و پاکسازی ورودی‌ها:**
  - اعتبارسنجی دقیق ورودی‌ها پیاده‌سازی شده تا از پرسش‌های متخاصم که می‌توانند آسیب‌پذیری‌های مدل را بهره‌برداری کنند، جلوگیری شود. اطمینان حاصل شود که خروجی‌ها اطلاعات حساس یا جزئیات داخلی مدل را فاش نکنند.
- **سامانه‌های تشخیص ناهنجاری:**
  - از ابزارهای تشخیص ناهنجاری برای شناسایی الگوهای غیرمعمول در استفاده از API، مانند افزایش ناگهانی حجم درخواست‌ها یا پرسش‌های تکراری استفاده شود. هشدارهایی برای فعالیت‌های مشکوک تنظیم شده تا پاسخ سریع به تلاش‌های احتمالی استخراج امکان‌پذیر باشد.
- **مدیریت پیچیدگی مدل:**
  - از تکنیک‌هایی استفاده شود که از بیش‌برازش مدل‌ها جلوگیری کند، زیرا این امر می‌تواند

پیش‌بینی‌پذیری و سهولت استخراج آن‌ها را افزایش دهد. هم‌چنین از روش‌های منظم‌سازی در حین آموزش استفاده شده تا پایداری مدل در برابر استخراج افزایش یابد.

- **رمزگذاری داده‌ها:**
  - از رمزگذاری برای داده‌های در حال استراحت و داده‌های در حال انتقال استفاده شود تا اطلاعات حساس استفاده‌شده توسط مدل‌های هوش مصنوعی محافظت شود. وزن‌ها و پارامترهای مدل رمزگذاری شده تا دسترسی غیرمجاز دشوارتر شود.
- **سیاست‌های کنترل دسترسی:**
  - اصل حداقل امتیاز پیاده‌سازی شود تا اطمینان حاصل شود که فقط افراد مجاز به مدل‌ها و داده‌های مربوطه دسترسی دارند. حسابرسی‌های منظم از لاگ‌های دسترسی و مجوزها انجام شده تا هرگونه دسترسی غیرمجاز شناسایی و اصلاح شود.
- **نظارت بر مدل:**
  - به‌طور مداوم عملکرد مدل نظارت شده تا ناهنجاری‌هایی که می‌توانند نشان‌دهنده دستکاری یا تلاش‌های استخراج باشند، شناسایی شوند. تغییرات در رفتار مدل تحلیل شود که می‌تواند نشانه‌ای از دسترسی غیرمجاز یا تکرار باشد.
- **اقدامات قانونی و رعایت مقررات:**
  - اطمینان حاصل شود که مدل‌ها و الگوریتم‌ها تحت پوشش محافظت‌های قانونی مناسب، مانند پتنت‌ها یا اسرار تجاری قرار دارند. پایبندی به قوانین و سازوکارهای مرتبط با داده تا عواقب قانونی ناشی از نقض داده‌ها کاهش یابد.
- **آموزش و آگاهی:**
  - دوره‌های آموزشی منظم برای کارکنان در مورد شناسایی و پاسخ به تهدیدات امنیتی مرتبط با مدل‌های هوش مصنوعی برگزار شود. برنامه‌های پاسخ به حوادث واضح و مشخصی برای مقابله با تلاش‌های احتمالی استخراج تدوین شده و به اشتراک گذاشته شود.
- **همکاری با کارشناسان امنیتی:**
  - با کارشناسان امنیت سایبری همکاری شده تا آسیب‌پذیری‌ها را ارزیابی کرده و بهترین شیوه‌ها برای ایمن‌سازی سامانه‌های هوش مصنوعی پیاده‌سازی شود. حملات شبیه‌سازی‌شده انجام شده تا نقاط ضعف در تدابیر امنیتی شناسایی و دفاع‌ها تقویت شوند.

۳-۳ بازسازی داده

حملات بازسازی داده[1] نوعی تهدید امنیتی هستند که شامل بازیابی یا بازسازی غیرمجاز داده‌های حساس یا خصوصی از اطلاعات یا خروجی‌های به ظاهر بی‌ضرر می‌شوند. این حملات از ضعف‌های موجود در مدیریت، ذخیره‌سازی و پردازش داده‌ها بهره‌برداری می‌کنند و معمولاً مدل‌های یادگیری ماشین، پایگاه‌های داده و سامانه‌های اشتراک‌گذاری داده را هدف قرار می‌دهند.

۳-۳-۱ ویژگی‌های کلیدی

حملات بازسازی داده دارای ویژگی‌های تعیین‌کننده‌ای هستند که سازوکارها، اهداف و تأثیرات آن‌ها را مشخص می‌کنند. درک این ویژگی‌ها برای شناسایی آسیب‌پذیری‌ها و پیاده‌سازی دفاع‌های مؤثر ضروری است. در اینجا ویژگی‌های کلیدی حملات بازسازی داده آورده شده است:

- **بهره‌برداری از داده‌های پردازش‌شده:** این حملات معمولاً داده‌هایی را هدف قرار می‌دهند که تغییر یافته، تجمیع شده یا ناشناس‌سازی شده‌اند. حمله‌کنندگان از ضعف‌های ذاتی در پردازش داده‌ها برای استنباط اطلاعات حساس اصلی بهره می‌برند. یک حمله‌کننده ممکن است خروجی‌های مدل را تجزیه و تحلیل کند تا نقاط داده‌ای آموزشی را بازسازی کند، حتی اگر داده‌های خام به‌طور مستقیم قابل دسترسی نباشند.
- **وابستگی به نشت داده:** حملات بازسازی داده معمولاً به نوعی نشت داده وابسته هستند، چه از طریق خروجی‌های مدل، متاداده یا ویژگی‌های آماری. خروجی‌های یک مدل یادگیری ماشین ممکن است الگوهایی را فاش کنند که می‌توانند برای استنباط ویژگی‌های خصوصی افراد مورد استفاده قرار گیرند.
- **استفاده از تکنیک‌های استنباط:** حمله‌کنندگان از تکنیک‌های مختلف استنباط، مانند عکس‌برداری مدل، برای بازسازی داده‌های حساس استفاده می‌کنند. این تکنیک‌ها رابطه بین ورودی‌ها و خروجی‌ها را برای استخراج اطلاعات پنهان تجزیه و تحلیل می‌کنند. در عکس‌برداری مدل، یک حمله‌کننده از خروجی‌های شناخته‌شده یک مدل یادگیری ماشین برای مهندسی معکوس ورودی‌ها استفاده می‌کند و ممکن است داده‌های حساس را فاش کند.
- **هدف‌گیری مدل‌های یادگیری ماشین:** بسیاری از حملات بازسازی داده به‌طور خاص مدل‌های یادگیری ماشین را هدف قرار می‌دهند که به‌دلیل وابستگی به مجموعه داده‌های بزرگ، به‌ویژه آسیب‌پذیر هستند. حمله‌کنندگان ممکن است از نمونه‌های متخاصم استفاده کنند تا مدل‌ها را فریب دهند و اطلاعاتی درباره داده‌های آموزشی آن‌ها فاش کنند.

1. Data Reconstruction

- پتانسیل برای دقت بالا: حملات بازسازی می‌توانند دقت بالایی در بازیابی داده‌های حساس به‌دست آورند، به‌ویژه هنگامی که اطلاعات کافی از خروجی‌های مدل یا منابع داده کمکی در دسترس باشد. حتی با داده‌های خروجی محدود، حمله‌کنندگان ممکن است با دقت شگفت‌آوری ورودی‌های اصلی را بازسازی کنند.
- **تأثیر بر حریم خصوصی و امنیت:** نگرانی اصلی در مورد حملات بازسازی داده، پتانسیل نقض حریم خصوصی و امنیت است که منجر به دسترسی غیرمجاز به اطلاعات حساس می‌شود. حملات موفق می‌توانند به سرقت هویت، به اشتراک‌گذاری غیرمجاز داده‌ها یا افشای اطلاعات تجاری محرمانه منجر شوند.
- **قابلیت کاربرد گسترده:** حملات بازسازی داده می‌توانند در حوزه‌های مختلفی از جمله بهداشت و درمان، مالی، رسانه‌های اجتماعی و غیره اعمال شوند و این امر آن‌ها را به تهدیدی چندمنظوره تبدیل می‌کند. در بخش بهداشت و درمان، داده‌های بازسازی‌شده بیماران می‌توانند منجر به نقض‌های شدید حریم خصوصی و عواقب نظارتی شوند.
- **تکنیک‌های تطبیقی:** حمله‌کنندگان می‌توانند راهبردهای خود را بر اساس سامانه خاص هدف‌گیری شده تطبیق دهند و از تلاش‌های قبلی برای بهبود نرخ موفقیت خود یاد بگیرند. اگر یک حمله به دلیل تدابیر خاصی ناموفق باشد، حمله‌کنندگان ممکن است رویکرد خود را برای دور زدن آن تدابیر در تلاش‌های آینده اصلاح کنند.
- **ریسک‌های قانونی و نظارتی:** سازمان‌هایی که قربانی حملات بازسازی داده می‌شوند، ممکن است با چالش‌های قانونی و نظارتی مواجه شوند، به‌ویژه در مورد قوانین و مقررات حفاظت از داده‌ها.
- **نیاز به دفاع‌های قوی:** دفاع‌های مؤثر در برابر حملات بازسازی داده نیازمند ترکیبی از تدابیر فنی، سیاسی و رویه‌ای برای حفاظت از اطلاعات حساس است. پیاده‌سازی شیوه‌هایی مانند حریم خصوصی تفاضلی، پروتکل‌های اشتراک‌گذاری داده امن و نظارت مستمر می‌تواند به کاهش خطرات کمک کند.

۲-۳-۳ دارایی‌های هدف

حملات بازسازی داده می‌توانند دارایی‌های متنوعی را هدف قرار دهند که هر کدام پیامدهای منحصر به فردی برای حریم خصوصی، امنیت و یکپارچگی عملیاتی دارند. در اینجا دارایی‌های اصلی که در معرض خطر حملات بازسازی داده هستند، آورده شده است:

- **مدل‌های یادگیری ماشین:** حمله‌کنندگان معمولاً مدل‌های یادگیری ماشین را هدف قرار می‌دهند تا داده‌های آموزشی حساس را استخراج کنند یا پارامترهای مدل را مهندسی معکوس

کنند. مدل‌های آسیب‌دیده می‌توانند به دسترسی غیرمجاز به الگوریتم‌های اختصاصی و اطلاعات حساس استفاده شده در حین آموزش منجر شوند.

- **پایگاه‌های داده:** پایگاه‌های داده ساختاریافته که شامل داده‌های شخصی، مالی یا پزشکی هستند، اهداف اصلی حملات بازسازی داده هستند. حملات موفق می‌توانند منجر به نقض داده‌ها، سرقت هویت و عدم تطابق با مقررات حفاظت از داده‌ها شوند.
- **خدمات ذخیره‌سازی ابری:** داده‌های ذخیره‌شده در محیط‌های ابری می‌توانند در صورت عدم تأمین مناسب، در معرض حملات بازسازی داده قرار گیرند. حمله‌کنندگان می‌توانند به اسناد حساس، داده‌های کاربران و اطلاعات اختصاصی دسترسی پیدا کنند که منجر به خسارات مالی و شهرت قابل توجهی می‌شود.
- **پروفایل‌های کاربری و داده‌های شخصی:** داده‌های شخصی، از جمله پروفایل‌های کاربران در پلتفرم‌های اجتماعی و سایت‌های تجارت الکترونیکی، می‌توانند از اطلاعات عمومی و نیمه‌عمومی بازسازی شوند. حمله‌کنندگان می‌توانند از داده‌های بازسازی‌شده برای فیشینگ، مهندسی اجتماعی یا سرقت هویت بهره‌برداری کنند.
- **سوابق بهداشتی:** سوابق پزشکی و داده‌های مرتبط با سلامتی بسیار حساس هستند و می‌توانند از منابع مختلفی، از جمله خروجی‌های مدل و داده‌های تجمیعی بازسازی شوند. نقض‌ها می‌توانند منجر به تخلفات جدی حریم خصوصی و عواقب قانونی تحت مقررات داده‌های بهداشتی احتمالی شوند.
- **داده‌های مالی:** تراکنش‌های مالی و اطلاعات حساب می‌توانند از الگوهای داده یا از طریق تعامل با مدل‌های مالی بازسازی شوند. آسیب‌پذیری داده‌های مالی می‌تواند منجر به تقلب، تراکنش‌های غیرمجاز و خسارات مالی قابل توجه برای افراد و سازمان‌ها شود.
- **داده‌های حسگر و اینترنت اشیا:** داده‌های تولید شده توسط دستگاه‌های اینترنت اشیا و حسگرها می‌توانند برای استنباط رفتار کاربران یا اطلاعات حساس عملیاتی بازسازی شوند. حمله‌کنندگان ممکن است به بینش‌هایی درباره عادات شخصی یا ضعف‌های عملیاتی در برنامه‌های صنعتی دست پیدا کنند که منجر به نقض امنیت می‌شود.
- **گزارش‌های هوش تجاری:** داده‌های تجمیع‌شده و گزارش‌های تحلیلی می‌توانند هدف قرار گیرند تا بینش‌ها یا راهبردهای حساس تجاری استخراج شوند. رقبا ممکن است از بینش‌های بازسازی‌شده برای به‌دست آوردن مزیت رقابتی یا تضعیف عملیات تجاری بهره‌برداری کنند.
- **مالکیت معنوی:** الگوریتم‌ها، مدل‌ها و داده‌های تحقیقاتی اختصاصی می‌توانند از خروجی‌ها یا تعاملات با سامانه‌ها بازسازی شوند. از دست دادن مالکیت معنوی می‌تواند مزیت رقابتی یک شرکت را تضعیف کرده و منجر به خسارات مالی شود.

- **داده‌های ارتباطی:** لاگ‌های ارتباطات (مانند ایمیل‌ها و پیام‌ها) می‌توانند برای استخراج اطلاعات حساس درباره افراد یا سازمان‌ها بازسازی شوند. دسترسی غیرمجاز به ارتباطات می‌تواند منجر به نقض اعتماد و محرمانگی شود.

۳-۳-۳ آسیب‌پذیری‌ها

حملات بازسازی داده از آسیب‌پذیری‌های خاصی در سامانه‌ها، فرآیندها و شیوه‌های مدیریت داده بهره‌برداری می‌کنند. درک این آسیب‌پذیری‌ها برای سازمان‌ها که به دنبال محافظت از اطلاعات حساس و کاهش خطرات مرتبط با چنین حملاتی هستند، ضروری است. در اینجا برخی از آسیب‌پذیری‌های رایج که می‌توانند در حملات بازسازی داده مورد بهره‌برداری قرار گیرند، آورده شده است:

- **ناقص بودن ناشناس‌سازی داده:** تکنیک‌های ناشناس‌سازی داده که به‌خوبی پیاده‌سازی نشده‌اند، می‌توانند ردیابی‌هایی را باقی بگذارند که حمله‌کنندگان از آن‌ها برای پیوند دادن داده‌های ناشناس به افراد استفاده کنند. برای مثال اگر شناسه‌هایی مانند تاریخ‌ها یا مکان‌ها به‌طور کافی حذف یا ناشناس نشوند، ممکن است مجوز شناسایی دوباره را فراهم کنند.
- **بیش‌برازش مدل:** مدل‌های یادگیری ماشین که بیش از حد برازش دارند، داده‌های آموزشی را حفظ می‌کنند و به‌جای تعمیم آن‌ها، کمتر در برابر حملات بازسازی مقاوم هستند. یک حمله‌کننده می‌تواند از خروجی‌های مدل برای استنباط نمونه‌های آموزشی خاص استفاده کند، به‌ویژه اگر مدل در آن ورودی‌ها عملکرد بسیار خوبی داشته باشد.
- **افشای خروجی‌های مدل:** فراهم کردن دسترسی به پیش‌بینی‌های مدل بدون محدودیت‌های کافی می‌تواند منجر به نشت اطلاعات درباره داده‌های زیرین شود. API های عمومی که پیش‌بینی‌های مدل را بازمی‌گردانند، می‌توانند به‌عنوان ابزاری برای جمع‌آوری اطلاعات درباره داده‌های آموزشی مورد استفاده قرار گیرند.
- **نشت داده از طریق اطلاعات کمکی:** حمله‌کنندگان ممکن است از داده‌های عمومی یا کمکی برای افزایش توانایی خود در بازسازی اطلاعات حساس از داده‌های پردازش‌شده استفاده کنند. ترکیب خروجی‌های مدل با مجموعه‌های داده خارجی می‌تواند دقت داده‌های بازسازی‌شده را بهبود بخشد.
- **حملات استنباط آماری:** حمله‌کنندگان می‌توانند از ویژگی‌های آماری مجموعه‌های داده برای استنباط ویژگی‌های حساس یا بازسازی نقاط داده اصلی استفاده کنند. برای مثال تجزیه و تحلیل توزیع خروجی‌های یک مدل پیش‌بینی ممکن است الگوهایی را فاش کند که با داده‌های ورودی خاص همبستگی دارد.
- **کنترل‌های دسترسی ناکافی:** کنترل‌های دسترسی ضعیف می‌توانند به کاربران غیرمجاز اجازه

دهند به سامانه‌ها، داده‌ها یا مدل‌های حساس دسترسی پیدا کنند و خطر حملات بازسازی را افزایش دهند. مجوزهای نامناسب پیکربندی شده در پایگاه‌های داده یا APIها می‌توانند داده‌های حساس را در معرض دسترسی بازیگران مخرب قرار دهند.

- **عدم امنیت در اشتراک‌گذاری داده:** سازمان‌هایی که داده‌ها را بدون تدابیر امنیتی مناسب به اشتراک می‌گذارند، ممکن است به‌طور ناخواسته اطلاعات حساس را افشا کنند که می‌توانند بازسازی شوند. به اشتراک‌گذاری داده‌های تجمیع‌شده بدون حفاظت کافی می‌تواند به حمله‌کنندگان اجازه دهد تا نقاط داده فردی را استنباط کنند.
- **خط لوله داده‌ها با طراحی ضعیف:** خط لوله‌های پردازش داده که فاقد تدابیر امنیتی هستند، می‌توانند نقاط آسیب‌پذیری شوند که اجازه بازسازی داده را می‌دهند. اگر تبدیل‌های داده شامل بررسی‌های امنیتی کافی نباشند، حمله‌کنندگان ممکن است از این ضعف‌ها بهره‌برداری کنند.
- **ذخیره‌سازی ناامن داده:** ذخیره‌سازی داده در محیط‌های ناامن، مانند پایگاه‌های داده یا خدمات ابری بدون رمزنگاری، می‌تواند آن را در معرض دسترسی غیرمجاز و بازسازی قرار دهد. یک حمله‌کننده که به پایگاه داده ابری ضعیف دسترسی پیدا کند، می‌تواند اطلاعات حساس را استخراج و تجزیه و تحلیل کند.
- **عدم اجرای حریم خصوصی تفاضلی:** عدم استفاده از تکنیک‌هایی مانند حریم خصوصی تفاضلی می‌تواند سامانه‌ها را در برابر حملات بهره‌برداری از خروجی‌های الگوریتم‌ها آسیب‌پذیر کند. بدون حریم خصوصی تفاضلی، خروجی‌های مدل‌های یادگیری ماشین ممکن است بیش از حد افشاگرانه باشند و به حمله‌کنندگان اجازه دهند داده‌های آموزشی را بازسازی کنند.

۳-۳-۴ سناریو تهدید

در یک حمله بازسازی داده، هدف مهاجم استخراج اطلاعات حساس از یک مدل آموزش‌دیده است. این می‌تواند زمانی رخ دهد که مدل‌ها به‌طور عمومی به اشتراک گذاشته می‌شوند، بینش‌ها از طریق APIها افشا می‌شوند، یا در فرآیندهای آموزشی مشترک مانند یادگیری فدرال اتفاق می‌افتد. مهاجم از خروجی‌های مدل، گرادیان‌ها یا حالت‌های داخلی برای مهندسی معکوس جنبه‌های داده‌های آموزشی استفاده می‌کند که این امر خطرات حریم خصوصی قابل توجهی را به‌ویژه اگر داده‌های حساسی مانند شناسایی‌های شخصی یا اطلاعات تجاری محرمانه درگیر باشند، ایجاد می‌کند.

- **دسترسی به مدل:**
  - مهاجم به مدل دسترسی پیدا می‌کند یا به‌طور مستقیم با به‌دست آوردن فایل‌های مدل،

یا به‌طور غیرمستقیم از طریق APIهای استنتاج که خروجی‌های مدل را افشا می‌کنند. این مرحله معمولاً شامل بهره‌برداری از پروتکل‌های امنیتی ضعیف یا استفاده از نقاط دسترسی قانونی است.

- **تحلیل خروجی‌ها یا گرادیان‌های مدل:**
  - مهاجم خروجی‌ها یا گرادیان‌های مدل (به‌عنوان مثال، از پاسخ‌ها به پرسش‌ها/ورودی‌های خاص) را بررسی می‌کند تا بینش‌هایی درباره ویژگی‌های داده جمع‌آوری کند. در مورد حریم خصوصی تفکیکی، هدف قرار دادن مدل‌ها در طول تبادل به‌روزرسانی‌ها یک روش رایج است.
- **انتخاب الگوریتم:**
  - مهاجم یک الگوریتم یا روش را برای استنتاج جزئیات از داده‌های جمع‌آوری‌شده انتخاب می‌کند. این ممکن است شامل تکنیک‌های بهینه‌سازی یا استفاده از آسیب‌پذیری‌های شناخته‌شده در معماری مدل باشد که آن‌ها را در معرض نشت داده قرار می‌دهد.
- **بازسازی داده‌های آموزشی:**
  - با استفاده از اطلاعات به‌دست‌آمده از مدل، مهاجم نمونه‌های آموزشی را بازسازی می‌کند. این می‌تواند شامل تکنیک‌هایی برای بازسازی نقاط داده دقیق یا شناسایی ویژگی‌های خاصی باشد که به قطعات خاصی از اطلاعات حساس اشاره می‌کند.
- **اعتبارسنجی و تکرار:**
  - مهاجم ممکن است بازسازی را در برابر داده‌های شناخته‌شده (اگر موجود باشد) اعتبارسنجی کند تا حمله را بیشتر تصحیح کرده و دقت نمونه‌های بازسازی‌شده را تضمین کند. این مرحله ممکن است شامل بهبود تکراری با استفاده از پرسش‌های اضافی یا خروجی‌های مدل باشد.

۵-۳-۳ نشانه‌های احتمال نفوذ

در اینجا برخی از نشانه‌های کلیدی که ممکن است نشان‌دهنده وقوع یا در حال وقوع یک حمله بازسازی داده باشند، آورده شده است:

- **پرسش‌های غیرمعمول مدل:** افزایش قابل توجه در تعداد یا فراوانی پرسش‌ها به یک مدل یادگیری ماشین، به‌ویژه با تغییراتی که به نظر می‌رسد برای استخراج اطلاعات خاص طراحی شده‌اند. نرخ بالای غیرطبیعی تماس‌های API از یک منبع خاص، به‌ویژه هنگام پرسش درباره ویژگی‌های حساس.

- **الگوهای دسترسی:** لاگ‌های دسترسی نشان‌دهنده الگوهای غیرمعمول، مانند درخواست‌هایی که در ساعات عجیب یا از آدرس‌های IP ناآشنا انجام می‌شوند. تلاش‌های مکرر برای دسترسی از یک منبع خارجی که بخشی از پایگاه کاربری معمول نیست.
- **پیش‌بینی‌های با اعتماد بالا:** ورود یک جریان از خروجی‌های مدل که نمرات اطمینان بالایی برای ورودی‌های خاص بازمی‌گرداند، که نشان می‌دهد ممکن است یک حمله‌کننده در حال کاوش در مدل باشد. درخواست‌های متعدد برای یک مجموعه خاص از نقاط داده که پیش‌بینی‌های ثابت و با اعتماد بالا تولید می‌کنند.
- **انحرافات داده:** تغییرات یا انحرافات غیرقابل توضیح در مجموعه‌های داده‌ای که برای آموزش یا ارزیابی استفاده می‌شوند، مانند نقاط داده غیرمنتظره یا تغییرات در توزیع داده. افزایش ناگهانی در ویژگی‌های خاص که قبلاً پایدار بودند.
- **دسترسی غیرمجاز به داده‌ها:** شناسایی تلاش‌های دسترسی غیرمجاز به مخازن داده حساس، از جمله پایگاه‌های داده یا ذخیره‌سازی ابری. لاگ‌های دسترسی که نشان‌دهنده تلاش‌های ورود ناموفق پس از آن با دسترسی موفق از حساب‌های ناشناخته است.
- **کاهش عملکرد مدل:** کاهش ناگهانی در عملکرد مدل که ممکن است نشان‌دهنده این باشد که مدل در حال بهره‌برداری یا دستکاری است. کاهش قابل توجه در دقت یا افزایش در طبقه‌بندی‌های نادرست.
- **رفتار مشکوک کاربران:** کاربران با رفتارهایی که از الگوهای عادی منحرف می‌شوند، مانند دسترسی به مقادیر زیادی از داده‌ها بدون هدف مشروع. یک کارمند مقدار زیادی از داده‌های حساس را دانلود می‌کند در حالی که با توجه به نقش او نیازی به آن‌ها ندارد.
- **گزارش‌های نشت داده:** گزارش‌ها یا هشدارهایی که نشان‌دهنده نشت داده‌های بالقوه هستند، چه از طریق حسابرسی‌های داخلی یا منابع خارجی و هشدارهایی از ابزارهای پیشگیری از نشت داده (DLP) که به اشتراک‌گذاری غیرمجاز داده‌ها را برجسته می‌کند.
- **تغییرات در شیوه‌های اشتراک‌گذاری داده:** تغییرات ناگهانی در شیوه‌های اشتراک‌گذاری داده، مانند همکاری‌های جدید یا اشتراک‌گذاری داده‌های حساس بدون تدابیر امنیتی مناسب. تصمیم ناگهانی برای اشتراک‌گذاری داده‌های تجمیع‌شده با طرف‌های ثالث بدون ناشناس‌سازی کافی.
- **افزایش همبستگی داده‌های خارجی:** یافته‌هایی که نشان می‌دهد مجموعه‌های داده خارجی در حال همبستگی با داده‌های داخلی هستند، که ممکن است نشان‌دهنده تلاش‌ها برای بازسازی اطلاعات حساس باشد. تجزیه و تحلیل مجموعه‌های داده عمومی که به‌طور نزدیک با مجموعه‌های داده داخلی هم‌راستا هستند و منجر به شناسایی دوباره می‌شود

۶-۳-۳ تأثیرات

در اینجا تأثیرات اصلی مرتبط با حملات بازسازی داده آورده شده است:

- **نقض حریم خصوصی:** افشای اطلاعات شخصی حساس می‌تواند منجر به نقض‌های قابل توجه حریم خصوصی برای افراد شود. قربانیان ممکن است دچار اضطراب، از دست دادن اعتماد و عواقب بلندمدت ناشی از سوءاستفاده از اطلاعات شخصی خود شوند.
- **زیان مالی:** سازمان‌ها ممکن است به دلیل حملات بازسازی داده متحمل زیان‌های مالی قابل توجهی شوند. هزینه‌ها می‌توانند ناشی از تلاش‌های ترمیم، هزینه‌های قانونی، جریمه‌های نظارتی و دعاوی احتمالی از سوی افراد یا نهادهای آسیب‌دیده باشند.
- **آسیب به شهرت:** نقض داده‌ها می‌تواند به شدت به شهرت سازمان آسیب بزند و بر اعتماد و وفاداری مشتریان تأثیر بگذارد و می‌تواند منجر به از دست دادن فرصت‌های تجاری، کاهش تعداد مشتریان و آسیب بلندمدت به ارزش برند شود.
- **عواقب قانونی:** سازمان‌ها ممکن است با جریمه‌های قانونی و نظارتی به دلیل عدم محافظت کافی از داده‌های حساس مواجه شوند.
- **اختلال در عملیات:** حملات بازسازی داده می‌توانند عملیات عادی کسب‌وکار را مختل کنند و منجر به زمان‌های غیرقابل دسترسی و کاهش بهره‌وری شوند. سازمان‌ها ممکن است نیاز به توقف عملیات، انجام تحقیقات و پیاده‌سازی تدابیر امنیتی جدید داشته باشند که همه این‌ها می‌تواند عملکرد را تحت تأثیر قرار دهد.
- **سرقت هویت:** داده‌های حساس بازسازی شده می‌توانند توسط حمله‌کنندگان برای ارتکاب سرقت هویت مورد استفاده قرار گیرند که منجر به عواقب شدید برای قربانیان می‌شود. قربانیان ممکن است با زیان مالی، آسیب به نمرات اعتباری و چالش‌های گسترده در بازیابی هویت خود روبه‌رو شوند.
- **از دست دادن مالکیت معنوی:** سازمان‌ها ممکن است به دلیل حملات بازسازی داده اطلاعات اختصاصی، الگوریتم‌ها یا اسرار تجاری خود را از دست بدهند. از دست دادن مالکیت معنوی می‌تواند موقعیت رقابتی شرکت را تضعیف کرده و منجر به عقب‌افتادگی در نوآوری شود.
- **افزایش هزینه‌های امنیتی:** پس از یک حمله بازسازی داده، سازمان‌ها معمولاً نیاز به سرمایه‌گذاری زیاد در تقویت امنیت و آموزش کارکنان دارند. افزایش بودجه‌های امنیت سایبری می‌تواند منابع را از سایر حوزه‌های بحرانی کسب‌وکار منحرف کند و بر رشد کلی تأثیر بگذارد.
- **کاهش اعتماد مشتری:** حوادث مکرر نقض داده‌ها می‌توانند منجر به کاهش قابل توجه اعتماد مشتریان به یک سازمان شوند. مشتریان ممکن است انتخاب کنند که کسب‌وکار

خود را به جاهای دیگر ببرند که منجر به عواقب مالی بلندمدت و از دست دادن سهم بازار می‌شود.

- **تأثیرات روانی:** افرادی که داده‌هایشان به خطر افتاده است ممکن است اضطراب، استرس و احساس آسیب‌پذیری را تجربه کنند. بار روانی می‌تواند قابل توجه باشد و بر رفاه قربانیان تأثیر بگذارد و منجر به ناراحتی عاطفی بلندمدت شود.

۳-۳-۷ راه‌های مقابله

رای مقابله مؤثر با حملات بازسازی داده، سازمان‌ها باید مجموعه‌ای از تدابیر فنی، مدیریتی و عملیاتی را پیاده‌سازی کنند. در اینجا راهبردهای کلیدی برای کاهش خطرات مرتبط با این حملات آورده شده است:

- **ناشناس‌سازی داده‌های قوی:** پیاده‌سازی تکنیک‌های ناشناس‌سازی قوی برای جلوگیری از شناسایی مجدد افراد در مجموعه‌های داده. استفاده از روش‌هایی مانند حریم خصوصی تفاضلی و اختلال داده‌ها برای حفاظت از اطلاعات حساس در حالی که کاربرد داده حفظ می‌شود.
- **کنترل دسترسی و احراز هویت:** اجرای کنترل‌های دسترسی دقیق برای محدود کردن اینکه چه کسانی می‌توانند به داده‌ها و مدل‌های حساس دسترسی پیدا کنند. پیاده‌سازی کنترل دسترسی مبتنی بر نقش (RBAC)[1]، احراز هویت چندعاملی و اصول حداقل دسترسی برای کاهش آسیب‌پذیری.
- **شیوه‌های امنیت مدل:** تأمین امنیت مدل‌های یادگیری ماشین در برابر حملات با اتخاذ بهترین شیوه‌ها در آموزش و استقرار مدل باید صورت پذیرد. انجام ممیزی‌های منظم برای شناسایی آسیب‌پذیری‌ها، اعمال آموزش ضدحمله و محدود کردن دسترسی مدل از طریق دروازه‌های API.
- **رمزگذاری داده:** رمزگذاری داده‌های حساس هم در حالت استراحت و هم در حین انتقال برای حفاظت از آن‌ها در برابر دسترسی غیرمجاز باید در نظر گرفته شود. استفاده از الگوریتم‌های رمزگذاری قوی و مدیریت کلیدهای رمزگذاری به‌طور ایمن برای تضمین محرمانگی داده.
- **ممیزی‌های امنیتی منظم:** انجام ارزیابی‌های امنیتی منظم برای شناسایی آسیب‌پذیری‌ها در سامانه‌ها و فرآیندها با بهره‌گیری از روش‌عایی مانند تست‌های نفوذ، اسکن آسیب‌پذیری و بررسی‌های امنیتی برای کشف نقاط ضعف قبل از اینکه مورد سوءاستفاده قرار گیرند.
- نظارت و تشخیص ناهنجاری: پیاده‌سازی سامانه‌های نظارتی برای شناسایی الگوهای

1. Role-based access control

غیرمعمول در دسترسی به داده و استفاده از مدل، ابزارهای تشخیص ناهنجاری و سازوکارهای لاگ‌گذاری برای شناسایی و پاسخ به فعالیت‌های مشکوک در زمان واقعی.

- **آموزش و آگاهی کاربران:** پرورش فرهنگ آگاهی امنیتی در میان کارکنان برای کاهش خطر تهدیدات داخلی و مهندسی اجتماعی و ارائه آموزش‌های منظم در مورد شیوه‌های امنیت داده، آگاهی از فیشینگ و رویه‌های گزارش‌دهی حوادث.
- **اصول حداقلی داده:** جمع‌آوری و نگهداری تنها حداقل مقدار داده‌های شخصی ضروری برای عملیات و مرور منظم شیوه‌های جمع‌آوری داده و حذف داده‌های غیرضروری برای کاهش آسیب‌پذیری در صورت نقض.
- **برنامه‌ریزی پاسخ به حوادث:** توسعه و نگهداری یک برنامه جامع پاسخ به حوادث برای رسیدگی مؤثر به نقض داده شامل مراحل شناسایی، قرنطینه، نابودی، بازیابی و ارتباطات برای کاهش تأثیر یک حمله.
- **همکاری و اشتراک‌گذاری اطلاعات:** همکاری با شرکای صنعتی و به اشتراک‌گذاری اطلاعات درباره تهدیدات و آسیب‌پذیری‌ها و شرکت در ابتکارات اشتراک‌گذاری اطلاعات تهدید و همکاری با سازمان‌های امنیت سایبری برای مطلع ماندن از تهدیدات نوظهور.

۳-۴ استنتاج عضویت

حملات استنتاج عضویت[1] نوعی حمله حریم خصوصی هستند که بر روی مدل‌های یادگیری ماشین متمرکز شده و در آن‌ها یک مهاجم سعی دارد تعیین کند که آیا یک داده خاص در مجموعه داده‌های آموزشی مدل گنجانده شده است یا نه. این نوع حمله نگرانی‌های قابل توجهی را در مورد حریم خصوصی افرادی که داده‌های آن‌ها ممکن است در آموزش مدل استفاده شده باشد، به ویژه در حوزه‌های حساس مانند مراقبت‌های بهداشتی، مالی و رسانه‌های اجتماعی، ایجاد می‌کند. حملات استنتاج عضویت نیاز به تکنیک‌های قوی حفظ حریم خصوصی در یادگیری ماشین را برجسته می‌کند. با ادامه رشد استفاده از مدل‌های یادگیری ماشین، درک و کاهش این خطرات برای محافظت از اطلاعات حساس و حفظ اعتماد کاربران ضروری است.

۳-۴-۱ ویژگی‌های کلیدی

حملات استنتاج عضویت ویژگی‌هایی دارند که نحوه عملکرد آن‌ها و پیامدهای مربوط به حریم خصوصی و امنیت در یادگیری ماشین را تعریف می‌کند. درک این ویژگی‌های کلیدی برای توسعه دفاع‌های مؤثر در برابر چنین حملاتی ضروری است. در اینجا ویژگی‌های اصلی حملات استنتاج عضویت آورده شده است:

1. Membership Inference

- **نقاط داده هدفمند:**
  - مهاجم بر روی نقاط داده خاص تمرکز می‌کند و سعی دارد تعیین کند که آیا آن‌ها در مجموعه داده‌های آموزشی گنجانده شده‌اند یا نه. این رویکرد هدفمند به مهاجمان اجازه می‌دهد تا بر روی افراد یا اطلاعات حساس متمرکز شوند و احتمال نقض حریم خصوصی را افزایش دهند.
- **خروجی‌های مدل:**
  - حمله بر اساس بررسی خروجی‌های مدل، به ویژه نمرات اطمینان مرتبط با پیش‌بینی‌ها، استوار است. نمرات اطمینان بالا برای ورودی‌های خاص اغلب نشان‌دهنده این است که آن‌ها بخشی از داده‌های آموزشی بوده‌اند، در حالی که نمرات پایین‌تر نشان می‌دهد که احتمالاً در آن‌ها گنجانده نشده‌اند.
- **دسترسی به مدل:**
  - مهاجمان معمولاً به نوعی دسترسی به مدل یادگیری ماشین نیاز دارند، چه از طریق API ها، اشتراک‌گذاری مدل یا سایر روش‌ها. درجه دسترسی می‌تواند متفاوت باشد، اما حتی دسترسی محدود نیز می‌تواند حملات مؤثر استنتاج عضویت را تسهیل کند.
- **تمرکز بر آموزش:**
  - حمله ممکن است از ویژگی‌های خاصی از فرآیند آموزش، مانند بیش‌برازش، بهره‌برداری کند، جایی که مدل بیش از حد از داده‌های آموزشی یاد می‌گیرد. مدل‌های بیش‌برازش شده تمایل دارند .نمرات اطمینان بالاتری برای نقاط داده آموزشی تولید کنند که آن‌ها را در برابر استنتاج عضویت آسیب‌پذیرتر می‌کند.
- **حساسیت داده:**
  - خطر و تأثیر حملات استنتاج عضویت در حوزه‌های دارای داده‌های حساس (مانند مراقبت‌های بهداشتی و مالی) افزایش می‌یابد. پتانسیل وجود نقض‌های شدید حریم خصوصی در این حوزه‌ها نیاز به تدابیر حفاظتی قوی‌تری را ایجاب می‌کند.
- **دانش مهاجمان:**
  - اثربخشی حمله ممکن است به دانش مهاجم درباره مدل و مجموعه داده بستگی داشته باشد. مهاجمان ممکن است از دانش قبلی درباره توزیع داده‌ها یا معماری‌های مدل برای افزایش شانس موفقیت خود استفاده کنند.
- **تنوع در موفقیت حمله:**
  - نرخ موفقیت حملات استنتاج عضویت می‌تواند بر اساس معماری مدل، مقدار داده‌های

آموزشی و راهبرد حمله خاص استفاده‌شده متفاوت باشد. این تنوع به این معنی است که برخی مدل‌ها ممکن است در برابر حملات مقاوم‌تر باشند، در حالی که دیگران ممکن است به طور قابل توجهی آسیب‌پذیرتر باشند.

- **آگاهی از تدابیر مقابله‌ای:**
  - مهاجمان ممکن است راهبردهای خود را بر اساس تدابیر مقابله‌ای که سازمان‌ها پیاده‌سازی می‌کنند (مانند حریم خصوصی تفاضلی) تطبیق دهند. تکامل مداوم هر دو راهبرد حمله و دفاع برای حفظ حریم خصوصی داده‌ها ضروری است.

۳-۴-۲ دارایی‌های هدف

حملات استنتاج عضویت می‌توانند دارایی‌های مختلفی را هدف قرار دهند که هر کدام پیامدهای متفاوتی برای حریم خصوصی و امنیت دارند. درک این دارایی‌های هدف برای سازمان‌ها ضروری است تا بتوانند تدابیر دفاعی مؤثری را پیاده‌سازی کنند. در اینجا دارایی‌های اصلی در معرض خطر از حملات استنتاج عضویت آورده شده است:

- **داده‌های آموزشی:**
  - مستقیم‌ترین هدف، مجموعه داده آموزشی است که برای توسعه مدل یادگیری ماشین استفاده می‌شود. اگر یک مهاجم بتواند تعیین کند که کدام نقاط داده در مجموعه آموزشی گنجانده شده‌اند، ممکن است اطلاعات شخصی حساس را افشا کند که منجر به نقض حریم خصوصی می‌شود.
- **حریم خصوصی کاربران:**
  - افرادی که داده‌های آن‌ها برای آموزش استفاده شده است، به ویژه در حوزه‌های حساس، در معرض خطر هستند. افشای هویت یا جزئیات شخصی کاربران می‌تواند منجر به سرقت هویت، تبعیض یا سایر پیامدهای مضری شود.
- **خروجی‌های مدل:**
  - خروجی‌های مدل یادگیری ماشین، از جمله پیش‌بینی‌ها و .نمرات اطمینان، نیز هدف قرار می‌گیرند. تجزیه و تحلیل این خروجی‌ها می‌تواند بینش‌هایی درباره اینکه آیا نقاط داده خاصی بخشی از داده‌های آموزشی بوده‌اند، فراهم کند و به نقض حریم خصوصی منجر شود.
- **مالکیت معنوی:**
  - الگوریتم‌ها و روش‌های پایه‌ای که برای توسعه مدل استفاده می‌شوند، در معرض خطر

هستند. اگر مهاجمان به بینش‌هایی درباره ویژگی‌های اختصاصی مدل یا روش‌های آموزشی دست یابند، ممکن است منجر به آسیب به رقابت یا از دست دادن اسرار تجاری شود.

- **اکوسامانه‌های داده:**
  - اکوسامانه داده گسترده‌تر، از جمله مشارکت‌ها و توافقات اشتراک‌گذاری داده، می‌تواند تحت تأثیر قرار گیرد. نقض‌ها ممکن است منجر به بی‌اعتمادی میان شرکا یا همکاران شوند که بر ابتکارات آینده اشتراک‌گذاری داده تأثیر می‌گذارد.

۳-۴-۳ آسیب‌پذیری‌ها

- حملات استنتاج عضویت از آسیب‌پذیری‌های خاصی در مدل‌های یادگیری ماشین و محیطهای استقرار آن‌ها بهره‌برداری می‌کنند. درک این آسیب‌پذیری‌ها برای توسعه تدابیر دفاعی مؤثر و افزایش امنیت داده‌ها ضروری است. در اینجا آسیب‌پذیری‌های اصلی که حملات استنتاج عضویت را ممکن می‌سازند، آورده شده است:

- **بیش‌برازش:**
  - مدل‌های بیش‌برازش معمولاً نمرات اطمینان بالاتری برای نقاط داده آموزشی تولید می‌کنند که این امر به مهاجمان اجازه می‌دهد به راحتی عضویت را استنتاج کنند.
- **معماری مدل:**
  - برخی از معماری‌های مدل به طور ذاتی اطلاعات بیشتری درباره داده‌های آموزشی افشا می‌کنند. مدل‌های پیچیده، مانند شبکه‌های عصبی عمیق، می‌توانند رفتارهای متمایزی را نشان دهند که مهاجمان می‌توانند از آن‌ها برای استنتاج عضویت استفاده کنند.
- **نمرات اطمینان خروجی:**
  - خروجی‌های مدل، به ویژه نمرات اطمینان مرتبط با پیش‌بینی‌ها، می‌توانند اطلاعاتی را درباره داده‌های آموزشی افشا کنند. .نمرات اطمینان بالا برای ورودی‌های خاص ممکن است نشان‌دهنده این باشد که آن ورودی‌ها بخشی از مجموعه آموزشی بوده‌اند که به تسهیل استنتاج عضویت منجر می‌شود.
- **کنترل‌های دسترسی ناکافی:**
  - کنترل‌های دسترسی ضعیف می‌توانند به کاربران غیرمجاز اجازه دهند تا مدل را پرسش‌گری کنند. اگر مهاجمان بتوانند به راحتی به مدل دسترسی پیدا کنند، می‌توانند بدون شناسایی به راحتی حملات استنتاج عضویت را انجام دهند.

- **عدم ناشناس‌سازی داده:**
  - عدم ناشناس‌سازی یا عدم شناسایی داده‌های آموزشی می‌تواند اطلاعات حساس را افشا کند. حتی اگر مهاجمان نتوانند به طور مستقیم به داده‌های آموزشی دسترسی پیدا کنند، ممکن است جزئیات مربوط به افراد را استنتاج کنند.
- **عدم تنظیم مناسب:**
  - تکنیک‌های تنظیم به جلوگیری از بیش‌براش کمک می‌کنند و مدل‌های پیچیده را جریمه می‌کنند. مدل‌هایی که تنظیم مناسبی ندارند، ممکن است بیشتر در برابر حملات استنتاج عضویت آسیب‌پذیر باشند، زیرا می‌توانند داده‌های آموزشی را به خاطر بسپارند.
- **افشای مدل:**
  - زمانی که مدل‌ها از طریق APIها یا محیط‌های مشترک در معرض دید قرار می‌گیرند، هدفی برای مهاجمان می‌شوند. مدل‌های قابل دسترسی عمومی می‌توانند به طور گسترده‌ای پرسش‌گری شوند که احتمال موفقیت استنتاج عضویت را افزایش می‌دهد.
- **دانش توزیع داده‌های آموزشی:**
  - مهاجمان ممکن است از دانش درباره توزیع داده‌های آموزشی بهره‌برداری کنند. اگر مهاجمان درک خوبی از انواع داده‌های استفاده‌شده برای آموزش داشته باشند، می‌توانند پرسش‌های مؤثرتری برای استنتاج عضویت طراحی کنند.
- **آگاهی محدود از خطرات حریم خصوصی:**
  - سازمان‌ها ممکن است خطرات مرتبط با حملات استنتاج عضویت را دست کم بگیرند. عدم آگاهی می‌تواند منجر به تدابیر امنیتی ناکافی و آسیب‌پذیری‌های بیشتر شود.
- **شیوه‌های بازآموزی مدل:**
  - عدم تناسب یا عدم نظم در بازآموزی مدل می‌تواند منجر به دفاع‌های قدیمی شود. اگر یک مدل به‌طور منظم به‌روزرسانی نشود، ممکن است آسیب‌پذیری‌هایی را حفظ کند که در نسخه‌های قبلی افشا شده‌اند و این امر موفقیت مهاجمان را تسهیل می‌کند.

۳-۴-۴ سناریو حمله

در یک حمله استنتاج عضویت، هدف مهاجم بهره‌برداری از پاسخ‌های مدل برای شناسایی این است که آیا نقاط داده خاصی بخشی از مجموعه داده‌های آموزشی آن بوده‌اند یا خیر. این تهدید به‌ویژه در سناریوهایی که مدل‌ها در زمینه‌های حساس مانند بهداشت و درمان یا سامانه‌های توصیه‌گری شخصی مستقر شده‌اند، نگران‌کننده است؛ جایی که افشای وجود داده‌های خاص می‌تواند منجر به

نقض حریم خصوصی یا مسائل اخلاقی شود.

- **دسترسی به مدل:**
  - مهاجم به مدل هدف دسترسی پیدا می‌کند، معمولاً از طریق APIهای عمومی که اجازه پرسش‌گری می‌دهند یا از طریق تعاملات مستقیم در محیط‌هایی که مدل در دسترس است.
- **پرسش از مدل:**
  - مهاجم پرسش‌های خاص یا ورودی‌های آزمایشی را به مدل ارسال می‌کند. این ورودی‌ها می‌توانند شامل نقاط داده‌ای باشند که مهاجم می‌خواهد برای عضویت در مجموعه داده آموزشی آزمایش کند.
- **تحلیل نمرات اطمینان مدل:**
  - برای هر پرسش، مهاجم خروجی مدل را مشاهده می‌کند و به‌ویژه بر .نمرات اطمینان یا توزیع‌های احتمالی تمرکز می‌کند. مدل‌ها معمولاً برای ورودی‌هایی که در طول آموزش با آن‌ها مواجه شده‌اند، اعتماد بالاتری نشان می‌دهند.
- **مقایسه با الگوهای شناخته‌شده:**
  - مهاجم پاسخ‌های مدل برای این پرسش‌ها را با پاسخ‌های مربوط به داده‌های آموزشی شناخته‌شده مقایسه می‌کند تا الگوها یا ناهنجاری‌هایی که نشان‌دهنده عضویت در مجموعه داده آموزشی است، شناسایی کند.
- **استنتاج وضعیت عضویت:**
  - مهاجم بر اساس پاسخ‌های مدل، درباره اینکه آیا هر ورودی آزمایش شده احتمالاً بخشی از مجموعه داده آموزشی بوده است یا خیر، تصمیم‌گیری می‌کند و از روش‌های آماری یا یادگیری ماشین برای بهبود دقت استفاده می‌کند.

۳-۴-۵ نشانه‌های احتمال نفوذ

در اینجا برخی از نشانه‌های کلیدی که ممکن است نشان‌دهنده وقوع یا در حال وقوع بودن یک حمله استنتاج عضویت باشند، آورده شده است:

- **پرسش‌های غیرمعمول مدل:**
  - افزایش قابل توجه در تعداد یا فراوانی پرسش‌ها به یک مدل یادگیری ماشین، به‌ویژه با تغییراتی که به نظر می‌رسد برای استخراج اطلاعات خاص طراحی شده‌اند. نرخ بالای غیرطبیعی تماس‌های API از یک منبع خاص، به‌ویژه هنگام پرسش درباره ویژگی‌های حساس.

- **الگوهای دسترسی:**
  - لاگ‌های دسترسی نشان‌دهنده الگوهای غیرمعمول، مانند درخواست‌هایی که در ساعات عجیب یا از آدرس‌های IP ناآشنا انجام می‌شوند. تلاش‌های مکرر برای دسترسی از یک منبع خارجی که بخشی از پایگاه کاربری معمول نیست.
- **پیش‌بینی‌های با اعتماد بالا:**
  - ورود یک جریان از خروجی‌های مدل که .نمرات اطمینان بالایی برای ورودی‌های خاص بازمی‌گرداند، که نشان می‌دهد ممکن است یک حمله‌کننده در حال کاوش در مدل باشد. درخواست‌های متعدد برای یک مجموعه خاص از نقاط داده که پیش‌بینی‌های ثابت و با اعتماد بالا تولید می‌کنند.
- **انحرافات داده:**
  - تغییرات یا انحرافات غیرقابل توضیح در مجموعه‌های داده‌ای که برای آموزش یا ارزیابی استفاده می‌شوند، مانند نقاط داده غیرمنتظره یا تغییرات در توزیع داده. افزایش ناگهانی در ویژگی‌های خاص که قبلاً پایدار بودند.
- **دسترسی غیرمجاز به داده‌ها:**
  - شناسایی تلاش‌های دسترسی غیرمجاز به مخازن داده حساس، از جمله پایگاه‌های داده یا ذخیره‌سازی ابری. لاگ‌های دسترسی که نشان‌دهنده تلاش‌های ورود ناموفق پس از آن با دسترسی موفق از حساب‌های ناشناخته است.
- **کاهش عملکرد مدل:**
  - کاهش ناگهانی در عملکرد مدل که ممکن است نشان‌دهنده این باشد که مدل در حال بهره‌برداری یا دستکاری است. کاهش قابل توجه در دقت یا افزایش در طبقه‌بندی‌های نادرست.
- **رفتار مشکوک کاربران:**
  - کاربران با رفتارهایی که از الگوهای عادی منحرف می‌شوند، مانند دسترسی به مقادیر زیادی از داده‌ها بدون هدف مشروع. یک کارمند دانلود مقدار زیادی از داده‌های حساس در حالی که نقش او نیاز به آن ندارد.
- **گزارش‌های نشت داده:**
  - گزارش‌ها یا هشدارهایی که نشان‌دهنده نشت داده‌های بالقوه هستند، چه از طریق حسابرسی‌های داخلی یا منابع خارجی. هشدارهایی از ابزارهای پیشگیری از نشت داده

(DLP)[1] که به اشتراک‌گذاری غیرمجاز داده‌ها را برجسته می‌کند.

- **تغییرات در شیوه‌های اشتراک‌گذاری داده:**
  - تغییرات ناگهانی در شیوه‌های اشتراک‌گذاری داده، مانند همکاری‌های جدید یا اشتراک‌گذاری داده‌های حساس بدون تدابیر امنیتی مناسب. تصمیم ناگهانی برای اشتراک‌گذاری داده‌های تجمیع‌شده با طرف‌های ثالث بدون ناشناس‌سازی کافی.
- **افزایش همبستگی داده‌های خارجی:**
  - یافته‌هایی که نشان می‌دهد مجموعه‌های داده خارجی در حال همبستگی با داده‌های داخلی هستند، که ممکن است نشان‌دهنده تلاش‌ها برای بازسازی اطلاعات حساس باشد. تجزیه و تحلیل مجموعه‌های داده عمومی که به‌طور نزدیک با مجموعه‌های داده داخلی هم‌راستا هستند و منجر به شناسایی دوباره می‌شود

۶-۴-۳ تأثیرات

در اینجا تأثیرات اصلی حملات استنتاج عضویت آورده شده است:

- **نقض حریم خصوصی:**
  - تأثیر فوری‌تر، نقض حریم خصوصی فردی است. افشای داده‌های شخصی از مجموعه‌های آموزشی می‌تواند به دسترسی غیرمجاز به اطلاعات حساس منجر شود و حریم خصوصی افراد را به خطر اندازد.
- **سرقت هویت:**
  - مهاجمان می‌توانند از اطلاعات استنتاج‌شده برای سرقت هویت استفاده کنند. این می‌تواند منجر به خسارت مالی برای قربانیان، آسیب به نمرات اعتباری و عواقب طولانی‌مدت بر زندگی شخصی و حرفه‌ای شود.
- **آسیب به اعتبار:**
  - سازمان‌ها زمانی که امنیت داده‌هایشان به خطر می‌افتد، آسیب به اعتبار می‌بینند. از دست دادن اعتماد مصرف‌کنندگان می‌تواند به کاهش وفاداری مشتری و تبلیغات منفی منجر شود و بر عملیات و درآمد کسب‌وکار تأثیر بگذارد.
- **عواقب قانونی و نظارتی:**
  - سازمان‌ها ممکن است با عواقب قانونی به خاطر عدم حفاظت از داده‌های حساس مواجه شوند. نقض مقررات حفاظت از داده‌ها می‌تواند منجر به جریمه‌ها، تحریم‌ها و افزایش

1. Data Loss Prevention

نظارت از سوی نهادهای نظارتی شود.

- **ضرر مالی:**
  - هزینه‌های مرتبط با نقض داده می‌تواند قابل توجه باشد. سازمان‌ها ممکن است هزینه‌هایی مربوط به پاسخ به حادثه، هزینه‌های قانونی، جریمه‌های نظارتی و از دست دادن کسب‌وکار به دلیل کاهش اعتماد مصرف‌کنندگان را متحمل شوند.
- **اختلال در عملیات:**
  - پاسخ به یک حمله استنتاج عضویت می‌تواند منابع را منحرف کرده و عملیات عادی را مختل کند. سازمان‌ها ممکن است نیاز به توقف خدمات، انجام تحقیقات و پیاده‌سازی تدابیر امنیتی جدید داشته باشند که منجر به زمان خاموشی و از دست رفتن بهره‌وری می‌شود.
- **سوءاستفاده‌های هدفمند:**
  - مهاجمان ممکن است از اطلاعات استنتاج‌شده برای حملات فیشینگ یا مهندسی اجتماعی هدفمند استفاده کنند. این می‌تواند به نقض‌های بیشتری منجر شود، زیرا افراد ممکن است نسبت به حملات متناسب با شرایط خاص خود آسیب‌پذیرتر باشند.
- **افزایش هزینه‌های امنیتی:**
  - سازمان‌ها ممکن است نیاز به سرمایه‌گذاری بیشتر در تدابیر امنیتی برای جلوگیری از حملات آینده داشته باشند. این می‌تواند به افزایش هزینه‌های عملیاتی منجر شود، زیرا سازمان‌ها تدابیر دفاعی و سامانه‌های نظارتی قوی‌تری را پیاده‌سازی می‌کنند.
- **کاهش اعتماد عمومی به سامانه‌های هوش مصنوعی:**
  - حملات استنتاج عضویت می‌توانند به بدبینی نسبت به ایمنی فناوری‌های هوش مصنوعی و یادگیری ماشین کمک کنند. اگر مصرف‌کنندگان احساس کنند که داده‌هایشان در معرض خطر است، ممکن است کمتر تمایل به تعامل با راه‌حل‌های هوش مصنوعی داشته باشند که مانع پیشرفت و پذیرش فناوری می‌شود.
- **تأثیرات منفی بر اشتراک‌گذاری داده‌ها و تحقیقات:**
  - نگرانی‌ها در مورد نقض حریم خصوصی می‌تواند سازمان‌ها را از اشتراک‌گذاری داده‌ها برای اهداف تحقیقاتی یا همکاری بازدارد. این می‌تواند نوآوری را کند کرده و پیشرفت‌ها در زمینه‌هایی که بر تحلیل داده‌های جمعی تکیه دارند، مانند بهداشت و علوم اجتماعی، را مختل کند.

۳-۴-۷ راه‌های مقابله با استنتاج عضویت

- برای محافظت در برابر حملات استنتاج عضویت، سازمان‌ها می‌توانند مجموعه‌ای از راهبردها و بهترین شیوه‌ها را پیاده‌سازی کنند. این تدابیر بر بهبود حریم خصوصی داده‌ها، تقویت امنیت مدل و افزایش تاب‌آوری کلی سامانه متمرکز هستند. در اینجا راه‌های کلیدی مقابله آورده شده است:

- **حریم خصوصی تفاضلی:**
  - گنجاندن حریم خصوصی تفاضلی در آموزش مدل می‌تواند به مبهم ساختن مشارکت‌های فردی کمک کند. به داده‌ها یا خروجی‌های مدل نویز اضافه کنید تا تشخیص اینکه آیا نقاط داده خاصی در مجموعه آموزشی گنجانده شده‌اند دشوار شود.
- **تکنیک‌های منظم‌سازی:**
  - از روش‌های منظم‌سازی برای جلوگیری از بیش‌برازش استفاده کنید که می‌تواند ویژگی‌های داده‌های آموزشی را افشا کند. تکنیک‌هایی مانند منظم‌سازی L1/L2، افت[1] یا توقف زودهنگام می‌توانند به حفظ تعمیم مدل کمک کنند.
- **ترکیب مدل‌ها:**
  - ترکیب چندین مدل می‌تواند احتمال استنتاج عضویت را کاهش دهد. روش‌های ترکیبی که پیش‌بینی‌ها را تجمیع می‌کنند تا مقاومت در برابر حملات استنتاج را افزایش دهند.
- **محدود کردن اطلاعات خروجی:**
  - کنترل دقت خروجی‌های مدل برای به حداقل رساندن افشای اطلاعات حساس. به جای ارائه پیش‌بینی‌های خام یا .نمرات اطمینان، خروجی‌ها را آستانه‌گذاری کرده یا فقط نتایج دسته‌ای ارائه شوند.
- **کنترل دسترسی و نظارت:**
  - پیاده‌سازی کنترل‌های دسترسی سختگیرانه و نظارت بر پرسش‌های مدل. از سازوکارهای احراز هویت و مجوز استفاده کرده تا دسترسی را محدود کرده و الگوهای استفاده را برای شناسایی ناهنجاری‌ها نظارت کنید.
- **اعتبارسنجی و پاکسازی ورودی:**
  - اطمینان از اعتبارسنجی و پاکسازی ورودی‌های مدل برای جلوگیری از بررسی‌های مخرب با پیاده‌سازی بررسی‌هایی که پرسش‌های مشکوک را فیلتر کرده یا محدود کردن انواع

1. Dropout

ورودی‌هایی که می‌توانند پردازش شوند.

- **ناشناس‌سازی داده‌ها:**
  - ناشناس‌سازی داده‌های آموزشی برای حفاظت از هویت‌های فردی و اطلاعات حساس با استفاده از تکنیک‌هایی مانند هش‌کردن، عمومی‌سازی برای مبهم‌سازی اطلاعات قابل شناسایی قبل از آموزش.
- **آموزش مدل قوی:**
  - آموزش مدل‌ها با استفاده از مجموعه‌های داده متنوع و نماینده برای کاهش بیش‌برازش و به روز رسانی مدل‌ها به‌طور منظم با داده‌های جدید به‌روزرسانی آموزش دوباره برای حفظ عملکرد و امنیت.
- **آموزش مقابله‌ای:**
  - گنجاندن نمونه‌های مقابله‌ای در طول آموزش برای افزایش تاب‌آوری مدل. مدل‌ها را با داده‌های اصلی و داده‌های مختل‌شده به‌صورت مقابله‌ای آموزش داده تا مقاومت در برابر حملات استنتاج افزایش یابد.
- **آموزش و آگاهی کاربران:**
  - آموزش کاربران درباره بهترین شیوه‌های حریم خصوصی و امنیت داده و برگزاری جلسات آموزشی یا فراهم کردن مواد آموزشی در مورد شناسایی و کاهش خطرات مرتبط با اشتراک‌گذاری و استفاده از داده‌ها.

۳-۵ استنتاج ویژگی

حملات استنتاج ویژگی[1] تهدیدی قابل‌توجه برای حریم خصوصی و امنیت مدل‌های یادگیری ماشین به شمار می‌روند. برخلاف حملات استنتاج عضویت که بر تعیین این‌که آیا یک نقطه داده خاص بخشی از مجموعه آموزشی بوده است یا خیر، متمرکز هستند، حملات استنتاج ویژگی هدفشان استخراج ویژگی‌ها یا خصوصیات عمومی درباره داده‌های آموزشی خود است. این می‌تواند شامل اطلاعات آماری حساس، جزئیات توزیع یا الگوهای کلی باشد که ممکن است بینش‌های محرمانه‌ای درباره افراد نمایانده شده در مجموعه داده‌های آموزشی افشا کند.

۳-۵-۱ ویژگی‌های کلیدی

حملات استنتاج ویژگی با چندین ویژگی متمایز مشخص می‌شوند که آن‌ها را از سایر انواع حملات، مانند استنتاج عضویت، متمایز می‌کند. درک این ویژگی‌های کلیدی برای شناسایی تأثیر بالقوه آن‌ها و

1. Property Inference

تدوین راهبردهای کاهش مؤثر ضروری است. در اینجا ویژگی‌های اصلی حملات استنتاج ویژگی آورده شده است:

- **تمرکز بر ویژگی‌های عمومی:**
  - بر خلاف حملات استنتاج عضویت که به دنبال شناسایی نقاط داده خاص در مجموعه آموزشی هستند، حملات استنتاج ویژگی هدفشان استنتاج ویژگی‌ها یا خصوصیات آماری عمومی داده‌های آموزشی است. این امر می‌تواند شامل استنتاج مقادیر متوسط، توزیع‌ها یا همبستگی‌ها بین ویژگی‌ها باشد.
- **وابستگی به مدل:**
  - برخی از معماری‌های مدل یا تکنیک‌های آموزشی ممکن است به حملات استنتاج ویژگی حساس‌تر باشند. مدل‌هایی که به داده‌های آموزشی بیش از حد تطابق پیدا می‌کنند یا در پیش‌بینی‌ها اعتماد بالایی نشان می‌دهند، می‌توانند اطلاعات بیشتری در مورد ویژگی‌های داده‌های زیرین افشا کنند.
- **استفاده از پاسخ‌های پرسش:**
  - مهاجمان از خروجی‌های مدل برای استنتاج ویژگی‌های داده‌های آموزشی استفاده می‌کنند. با پرسش‌های نظام‌مند از مدل با ورودی‌های مختلف و تجزیه و تحلیل پاسخ‌ها، مهاجمان می‌توانند اطلاعات ارزشمندی در مورد توزیع‌های داده استخراج کنند.
- **پتانسیل بازسازی داده:**
  - در برخی موارد، حملات استنتاج ویژگی می‌توانند به بازسازی داده‌های حساس منجر شوند، به ویژه اگر مدل بیش از حد اطلاعاتی باشد. مهاجمان ممکن است جنبه‌هایی از مجموعه داده‌های آموزشی را بر اساس ویژگی‌های آماری استنتاج شده بازسازی کنند.
- **تهدیدی برای حریم خصوصی و محرمانگی:**
  - حملات استنتاج ویژگی می‌توانند اطلاعات حساس را افشا کنند و به نقض حریم خصوصی منجر شوند. این می‌تواند عواقب جدی در حوزه‌هایی مانند بهداشت و درمان، جایی که محرمانگی داده‌های بیماران بسیار حائز اهمیت است، داشته باشد.
- **تأثیر بر کارایی مدل:**
  - تکنیک‌های استفاده شده برای کاهش حملات استنتاج ویژگی می‌توانند به طور ناخواسته بر عملکرد و کارایی مدل تأثیر بگذارند. پیاده‌سازی تدابیر حریم خصوصی قوی ممکن است دقت یا اثربخشی مدل را در وظایف مورد نظر کاهش دهد.

- **نیاز به راهبردهای هدفمند:**
  - کاهش حملات استنتاج ویژگی نیاز به راهبردهای خاصی دارد که متناسب با نوع داده و مدل مورد استفاده باشد. تکنیک‌هایی مانند حریم خصوصی تفاضلی، آموزش مقاوم و کنترل دسترسی باید با توجه به معماری و کاربرد مدل به دقت در نظر گرفته شوند.
- **تکنیک‌های حمله در حال تحول:**
  - با پیشرفت فناوری‌های دفاعی، روش‌های مورد استفاده توسط مهاجمان نیز بهبود می‌یابد. مهاجمان ممکن است راهبردهای پرسش جدیدی توسعه دهند یا از پیشرفت‌های یادگیری ماشین برای افزایش اثربخشی حملات استنتاج ویژگی استفاده کنند.
- **پیامدهای گسترده‌تر برای اشتراک‌گذاری داده:**
  - حملات استنتاج ویژگی می‌توانند از اشتراک‌گذاری داده و تلاش‌های تحقیقاتی همکاری جلوگیری کنند، زیرا سازمان‌ها ممکن است از افشای اطلاعات حساس بترسند. این می‌تواند نوآوری را کند کرده و پیشرفت‌ها در زمینه‌هایی که به داده‌های مشترک وابسته‌اند، مانند بهداشت و درمان و تحقیقات علوم اجتماعی، محدود کند.

۳-۵-۲ دارایی‌های هدف

حملات استنتاج ویژگی می‌توانند دارایی‌های مختلفی را در یک سازمان هدف قرار دهند، به‌ویژه آن‌هایی که به مدل‌های یادگیری ماشین و داده‌های مورد استفاده آن‌ها مربوط می‌شوند. در اینجا دارایی‌های اصلی در معرض خطر ناشی از حملات استنتاج ویژگی آورده شده است:

- **داده‌های آموزشی:**
  - مجموعه داده‌ای که برای آموزش مدل‌های یادگیری ماشین استفاده می‌شود، هدف اصلی است. مهاجمان ممکن است ویژگی‌ها یا آمار عمومی درباره داده‌های آموزشی استخراج کنند که می‌تواند به نقض حریم خصوصی و افشای اطلاعات حساس منجر شود.
- **پارامترهای مدل:**
  - پارامترها و وزن‌های داخلی مدل‌های یادگیری ماشین می‌توانند بینش‌هایی درباره داده‌های آموزشی ارائه دهند. آگاهی از این پارامترها ممکن است به مهاجمان اجازه دهد ویژگی‌هایی مانند توزیع داده‌ها یا همبستگی بین ویژگی‌ها را استنتاج کنند.
- **خروجی‌های مدل:**
  - پیش‌بینی‌ها و .نمرات اطمینان تولید شده توسط مدل، اطلاعات ارزشمندی برای مهاجمان به شمار می‌روند. تجزیه و تحلیل خروجی‌ها می‌تواند به مهاجمان کمک کند تا ویژگی‌های

آماری داده‌های آموزشی را استنتاج کنند که ممکن است برای مقاصد مخرب مورد استفاده قرار گیرد.

- **اطلاعات کاربران:**
  - اطلاعات مربوط به کاربران یا افرادی که در داده‌های آموزشی نمایانده شده‌اند می‌تواند هدف قرار گیرد. اگر مهاجمان بتوانند ویژگی‌های حساس درباره کاربران را استنتاج کنند، ممکن است به پروفایل‌سازی غیرمجاز یا حملات فیشینگ هدفمند منجر شود.
- **اطلاعات تجاری:**
  - بینش‌های حاصل از پیش‌بینی‌های مدل می‌تواند مزایای رقابتی فراهم کند. مهاجمان ممکن است از ویژگی‌های استنتاج شده داده‌ها برای به دست آوردن بینش‌هایی در مورد روندهای بازار، رفتار مشتری یا کارایی‌های عملیاتی استفاده کنند.
- **مالکیت معنوی:**
  - تکنیک‌ها، الگوریتم‌ها یا مجموعه داده‌هایی که توسط سازمان توسعه یافته‌اند ممکن است در معرض خطر باشند. اگر مهاجمان ویژگی‌های داده‌های مالکیتی را استنتاج کنند، می‌تواند به تضعیف مزیت رقابتی سازمان و احتمال سرقت مالکیت معنوی منجر شود.
- **تحقیق و توسعه:**
  - تلاش‌های تحقیقاتی جاری که به داده‌های حساس وابسته‌اند ممکن است به خطر بیفتند. حملات استنتاج ویژگی ممکن است همکاری یا اشتراک داده‌ها را بین محققان متوقف کرده و نوآوری و پیشرفت در زمینه‌های مختلف را محدود کنند.

۳-۵-۳ آسیب‌پذیری‌ها

در اینجا آسیب‌پذیری‌های کلیدی که می‌توانند هدف حملات استنتاج ویژگی قرار گیرند، آورده شده است:

- **بیش‌برازش:**
  - زمانی که یک مدل داده‌های آموزشی را به خوبی یاد می‌گیرد، ممکن است به نقاط داده خاصی بیش از حد حساس شود. مدل‌های تطابق یافته ممکن است به طور ناخواسته ویژگی‌های دقیقی از داده‌های آموزشی را افشا کنند و به مهاجمان اجازه دهند تا ویژگی‌های آماری را استنتاج کنند.
- **شفافیت بیش از حد مدل:**
  - مدل‌هایی که به راحتی قابل تفسیر هستند یا خروجی‌های دقیقی ارائه می‌دهند، می‌توانند

اطلاعات حساس را افشا کنند. اگر مهاجمان به پیش‌بینی‌های خام یا .نمرات اطمینان دسترسی پیدا کنند، می‌توانند این خروجی‌ها را برای استخراج بینش‌هایی در مورد ویژگی‌های داده‌های آموزشی تجزیه و تحلیل کنند.

- **عدم کافی بودن ناشناس‌سازی داده:**
  - اگر داده‌های آموزشی به درستی ناشناس یا تجمیع نشوند، اطلاعات حساس ممکن است قابل شناسایی باقی بمانند. مهاجمان ممکن است از تکنیک‌های ناشناس‌سازی ضعیف بهره‌برداری کنند تا ویژگی‌های خاصی درباره افراد در مجموعه داده‌ها استنتاج کنند.
- **پیچیدگی بالای مدل:**
  - مدل‌های پیچیده، مانند معماری‌های یادگیری عمیق، می‌توانند الگوهای پیچیده‌ای را در داده‌ها شناسایی کنند. این پیچیدگی می‌تواند به مهاجمان اجازه دهد تا ویژگی‌های دقیقی را استنتاج کنند، زیرا مدل ممکن است اطلاعات بیشتری درباره داده‌های آموزشی حفظ کند.
- **محدودیت‌های ناکافی در پرسش‌ها:**
  - عدم وجود کنترل بر اینکه چند پرسش و از چه نوعی می‌توان به مدل وارد کرد، می‌تواند مورد سوءاستفاده قرار گیرد. مهاجمان ممکن است به طور گسترده‌ای اطلاعات را جمع‌آوری کنند و احتمال موفقیت در استنتاج ویژگی‌های داده‌ها را افزایش دهند.
- **تعمیم‌پذیری ناکافی مدل:**
  - مدل‌هایی که به خوبی عمومیت پیدا نمی‌کنند، ممکن است ویژگی‌های خاصی از داده‌های آموزشی را افشا کنند. عمومیت ناکافی می‌تواند به آسیب‌پذیری‌هایی منجر شود که در آن‌ها مهاجمان می‌توانند بینش‌هایی درباره توزیع داده‌های زیرین به دست آورند.
- **نشت داده در حین آموزش:**
  - افشای غیر عمدی اطلاعات حساس در طول فرآیند آموزش می‌تواند آسیب‌پذیری‌هایی ایجاد کند. اگر داده‌ها در حین آموزش نشت یا به درستی مدیریت نشوند، خطر استنتاج ویژگی‌ها افزایش می‌یابد.
- **عدم وجود حریم خصوصی تفاضلی:**
  - بدون پیاده‌سازی تدابیر حریم خصوصی تفاضلی، مدل‌ها ممکن است اطلاعات زیادی درباره داده‌های آموزشی ارائه دهند. این عدم وجود تضمین‌های حریم خصوصی می‌تواند به مهاجمان اجازه دهد تا ویژگی‌های حساس را از خروجی‌های مدل استنتاج کنند.

- **شیوه‌های نامنظم مدیریت داده:**
  - مدیریت ضعیف داده و حاکمیت می‌تواند به آسیب‌پذیری‌ها در پردازش و ذخیره‌سازی داده‌ها منجر شود. شیوه‌های نامنظم می‌توانند منجر به افشای داده‌های حساس شوند که باید محافظت می‌شدند و به مهاجمان اجازه می‌دهند تا ویژگی‌ها را استنتاج کنند.
- **کنترل‌های امنیتی ناکافی:**
  - کنترل‌های دسترسی و نظارت ضعیف بر استفاده از مدل می‌تواند نقاط ورود برای مهاجمان ایجاد کند. اگر مهاجمان بتوانند بدون محدودیت به مدل‌ها دسترسی پیدا کنند، می‌توانند از آن‌ها برای جمع‌آوری اطلاعات درباره ویژگی‌های داده‌های آموزشی بهره‌برداری کنند.

۳-۵-۴ سناریو تهدید

- در یک حمله استنتاج ویژگی، مهاجم سعی دارد ویژگی‌های سطحی داده‌های آموزشی را با بهره‌برداری از خروجی‌ها یا ویژگی‌های یک مدل یادگیری ماشین استنتاج کند. این تهدید به‌ویژه در سناریوهایی نگران‌کننده است که داده‌های آموزشی شامل ویژگی‌های حساسی هستند که باید محرمانه بمانند، مانند اطلاعات جمعیتی یا الگوهای استفاده.
- **دسترسی به مدل:**
  - مهاجم به مدل دسترسی پیدا می‌کند، معمولاً از طریق APIهای عمومی، خدمات مستقر یا تعامل مستقیم با سامانه‌هایی که مدل در آن‌ها عملیاتی است.
- **پرسش از مدل:**
  - مهاجم با ورودی‌هایی که از توزیع‌های شناخته‌شده نمونه‌برداری شده‌اند یا به‌طور راهبردک انتخاب شده‌اند تا حداکثر اطلاعات را درباره ویژگی‌های مجموعه داده به‌دست آورد، از مدل پرسش می‌کند. این شامل استفاده از دامنه‌ای از تغییرات ورودی برای تحلیل رفتار مدل به‌طور جامع است.
- **مشاهده و جمع‌آوری داده:**
  - مهاجم به‌دقت خروجی‌های مدل را برای این ورودی‌ها مشاهده می‌کند و بر تغییرات پیش‌بینی‌ها، نمرات اعتماد یا ویژگی‌های دیگر خروجی که ممکن است بینش‌هایی درباره ویژگی‌های زیرین مجموعه داده ارائه دهد، متمرکز می‌شود.
- **مقایسه تحلیلی:**
  - با استفاده از تحلیل آماری، مهاجم خروجی‌های مشاهده‌شده مدل را با پاسخ‌های پیش‌بینی‌شده از مدل‌های آموزش‌دیده بر روی مجموعه داده‌های پایه با ویژگی‌های شناخته‌شده مقایسه

می‌کند. این کمک می‌کند تا رفتارهای منحصر به فردی که به ویژگی‌های خاص مرتبط است شناسایی شود.

- **استنتاج ویژگی‌ها:**
  - مهاجم از داده‌های جمع‌آوری‌شده برای استنتاج ویژگی‌های آماری مجموعه داده آموزشی استفاده می‌کند. این ممکن است شامل اعمال تکنیک‌های یادگیری ماشین باشد که ویژگی‌های مشاهده‌شده مدل را با ویژگی‌های خاص مجموعه داده مرتبط می‌کند.

۳-۵-۵ نشانه‌های احتمال نفوذ

در اینجا نشانه‌های کلیدی مرتبط با حملات استنتاج ویژگی آورده شده است:

- **الگوهای نامعمول پرسش:**
  - افزایش ناگهانی در حجم یا فراوانی پرسش‌ها به مدل یادگیری ماشین از طریق پرسش‌های مکرر یا نظام‌مند از ورودی‌های خاص که هدف آن‌ها استخراج ویژگی‌های خاص از مدل است.
- **تحلیل پاسخ‌های غیرعادی:**
  - پاسخ‌های مدل که از خروجی‌ها یا الگوهای مورد انتظار منحرف می‌شوند. .نمرات اطمینان یا پیش‌بینی‌های غیرمنتظره که نشان می‌دهد مهاجم در حال جستجوی جزئیات ویژگی‌های داده است.
- **لاگ‌های دسترسی با فعالیت مشکوک:**
  - الگوهای دسترسی غیرعادی در لاگ‌ها که نشان‌دهنده دسترسی غیرمجاز یا بیش از حد به مدل یا خروجی‌های آن هستند. تلاش‌های دسترسی از آدرس‌های IP نامشخص، مکان‌های جغرافیایی غیرمعمول یا در ساعات غیرمعمول.
- **تنوع بالای پرسش‌ها:**
  - دامنه وسیعی از ورودی‌های پرسش که به‌ویژه به نظر می‌رسد برای پوشش دادن یک طیف وسیع از ویژگی‌های داده طراحی شده‌اند. پرسش‌هایی که به‌طور راهبردی طراحی شده‌اند تا اطلاعات آماری خاصی درباره داده‌های آموزشی استخراج کنند.
- **آزمایش مکرر موارد مرزی:**
  - آزمایش مکرر موارد مرزی یا شرایط خاص که ممکن است جزئیات بیشتری از رفتار مدل را افشا کند. الگوی پرسش‌هایی که به سناریوهای کمتر رایج هدف‌گذاری شده‌اند و می‌توانند اطلاعات آماری حساس به‌دست آورند.

- **افزایش درخواست‌های خروجی مدل:**
  - افزایش درخواست‌ها برای خروجی‌های مدل که پیش‌بینی‌ها یا .نمرات اطمینان دقیقی را ارائه می‌دهند. درخواست‌هایی برای خروجی که برای استفاده معمول مدل غیرمعمول هستند و نشان‌دهنده رفتار جستجوگرانه‌اند.
- **تغییرات در عملکرد مدل:**
  - کاهش قابل‌توجه در عملکرد یا دقت مدل پس از الگوی پرسش غیرمعمول. افت توانایی‌های عمومی‌سازی که ممکن است نشان‌دهنده سوءاستفاده از مدل برای استخراج اطلاعات حساس باشد.
- **ناهنجاری‌های رفتار کاربر:**
  - تغییرات در الگوهای رفتار کاربر که ممکن است نشان‌دهنده تهدیدات داخلی یا حملات خارجی باشد. دسترسی کاربران به داده‌ها یا مدل‌ها خارج از دامنه عادی کار یا تخصص آن‌ها.
- **درخواست‌های خروجی غیرمعمول داده:**
  - درخواست‌هایی برای خروجی‌هایی که یا خیلی دقیق هستند یا با موارد استفاده مشروع همخوانی ندارند. درخواست‌هایی برای پیش‌بینی‌های بسیار جزئی که نشان‌دهنده قصد استنتاج ویژگی‌های داده‌های آموزشی است.
- **حلقه‌های بازخورد در پرسش‌های مدل:**
  - الگوهای پرسش که در آن پاسخ‌ها برای اطلاع‌رسانی به پرسش‌های بعدی استفاده می‌شوند. فرآیند تکراری که در آن مهاجم پرسش‌های خود را بر اساس خروجی‌های قبلی اصلاح می‌کند و نشان‌دهنده یک رویکرد نظام‌مند به استنتاج ویژگی‌ها است.

۳-۵-۶ تأثیرات استنتاج ویژگی

در اینجا تأثیرات کلیدی مرتبط با حملات استنتاج ویژگی آورده شده است:

- **نقض حریم خصوصی داده‌ها:**
  - حملات استنتاج می‌توانند منجر به افشای غیرمجاز داده‌های حساس و شخصی شوند. نقض حریم خصوصی ممکن است به عواقب قانونی و از دست دادن اعتماد مشتریان و ذینفعان منجر شود.
- **زیان مالی:**
  - سازمان‌ها ممکن است به دلیل تقلب یا هزینه‌های مربوط به نقض داده‌ها، زیان مالی

مستقیم متحمل شوند. این موضوع می‌تواند شامل هزینه‌های تعمیرات، حق‌الزحمه‌های قانونی، جریمه‌های نظارتی و از دست دادن فرصت‌های تجاری باشد.

- **آسیب به شهرت:**
  - حملات موفق استنتاج ویژگی می‌توانند شهرت سازمان‌ها را خدشه‌دار کنند. از دست دادن اعتماد مشتری و پوشش رسانه‌ای منفی می‌تواند بر چشم‌اندازهای تجاری و مشارکت‌های آینده تأثیر منفی بگذارد.
- **عواقب نظارتی:**
  - سازمان‌ها ممکن است به دلیل عدم حفاظت کافی از داده‌های حساس با جریمه‌ها مواجه شوند. نهادهای نظارتی ممکن است جریمه‌ها و تحریم‌هایی را به ویژه در صنایع تحت قوانین سختگیرانه حفاظت از داده‌ها اعمال کنند.
- **عدم برتری رقابتی:**
  - رقبا ممکن است از طریق حملات استنتاج ویژگی به بینش‌هایی درباره داده‌های اختصاصی، راهبردها یا رفتار مشتریان دست یابند. این می‌تواند موقعیت رقابتی سازمان را در بازار تضعیف کند و به از دست دادن فرصت‌ها و کاهش سهم بازار منجر شود.
- **اختلال در عملیات:**
  - حملات استنتاج ویژگی می‌توانند عملیات عادی کسب‌وکار را مختل کنند. سازمان‌ها ممکن است نیاز داشته باشند منابع خود را برای رسیدگی به نقض‌های امنیتی معطوف کنند که بر تولید و کارایی عملیاتی تأثیر منفی می‌گذارد.
- **افزایش هزینه‌های امنیتی:**
  - پس از حمله استنتاج ویژگی، سازمان‌ها ممکن است نیاز به سرمایه‌گذاری سنگین در تدابیر امنیتی پیشرفته داشته باشند. این موضوع شامل پیاده‌سازی فناوری‌های پیشرفته حفاظت از داده، آموزش کارکنان و انجام ممیزی‌های امنیتی منظم است.
- **از دست دادن مالکیت معنوی:**
  - مهاجمان می‌توانند ویژگی‌های حساس مرتبط با الگوریتم‌های اختصاصی، مدل‌ها یا راهبردهای تجاری را استنتاج کنند. این مورد می‌تواند منجر به از دست دادن برتری رقابتی و تضعیف تلاش‌های نوآوری شود.
- **نگرانی‌های اخلاقی:**
  - حملات استنتاج ویژگی مسائل اخلاقی مربوط به استفاده از داده‌ها و رضایت را به وجود

می‌آورند. سازمان‌ها ممکن است با واکنش عمومی مواجه شوند به دلیل ناکامی در حفاظت از حقوق داده‌های افراد، که منجر به مشکلات و بررسی‌های اخلاقی می‌شود.

- **کاهش اعتماد در درازمدت:**
  - حوادث مکرر افشای داده‌ها می‌تواند به آسیب قابل‌توجهی به اعتماد به فناوری‌های مبتنی بر داده منجر شود. این امر می‌تواند به پذیرش کندتر راه‌حل‌ها و فناوری‌های نوآورانه منجر شود زیرا مصرف‌کنندگان در استفاده از داده‌های خود محتاط می‌شوند.

۳-۵-۷ راه‌های مقابله

در اینجا تدابیر کلیدی برای مقابله با حملات استنتاج ویژگی آورده شده است:

- **پیاده‌سازی حریم خصوصی تفاضلی:**
  - از تکنیک‌های حریم خصوصی تفاضلی برای افزودن نویز به خروجی‌های مدل استفاده کنید تا برای مهاجمان دشوار شود که ویژگی‌های خاص داده‌های آموزشی را استنتاج کنند. این تکنیک‌ها به حفاظت از نقاط داده فردی کمک می‌کند در حالی که بینش‌های تجمعی مفیدی را ارائه می‌دهد.
- **محدود کردن دسترسی و قابلیت‌های پرسش:**
  - دسترسی به مدل‌های یادگیری ماشین را محدود کرده و انواع و فراوانی پرسش‌ها را کاهش دهید. این خطر بررسی مدل توسط مهاجمان برای اطلاعات حساس را کاهش می‌دهد.
- **استفاده از تقطیر مدل:**
  - یک مدل ساده‌تر را برای تقریب رفتار یک مدل پیچیده آموزش دهید در حالی که نشت اطلاعات را کاهش می‌دهد. تقطیر می‌تواند به پنهان‌سازی ویژگی‌های داده‌های زیرین کمک کند و خطر حملات استنتاج را کاهش دهد.
- **ممیزی و نظارت منظم:**
  - ممیزی‌های منظم از لاگ‌های دسترسی به مدل و الگوهای پرسش برای شناسایی فعالیت‌های مشکوک انجام دهید. شناسایی زودهنگام دسترسی‌های غیرمعمول می‌تواند به کاهش حملات بالقوه قبل از وقوع آسیب‌های جدی کمک کند.
- **استفاده از تکنیک‌های آموزش مدل مقاوم:**
  - از تکنیک‌هایی مانند آموزش خصمانه برای بهبود مقاومت مدل در برابر حملات استنتاج استفاده کنید. مدل‌های مقاوم کمتر احتمال دارد که اطلاعات حساس را از طریق خروجی‌های خود افشا کنند.

- **ناشناس‌سازی و تجمیع داده‌ها:**
  - اطمینان حاصل کنید که داده‌های حساس قبل از استفاده برای آموزش، ناشناس‌سازی یا تجمیع شده‌اند. این احتمال شناسایی مجدد را کاهش می‌دهد و از حریم خصوصی فردی محافظت می‌کند.
- **استفاده از کنترل‌های دسترسی و احراز هویت:**
  - کنترل‌های دسترسی و سازوکارهای احراز هویت سخت‌گیرانه‌ای پیاده‌سازی کنید تا فقط کاربران مجاز به داده‌ها و مدل‌های حساس دسترسی داشته باشند. این فرصت‌های سوءاستفاده از سوی کارمندان بدخواه یا مهاجمان خارجی را محدود می‌کند.
- **آموزش کارکنان و ذینفعان:**
  - آموزش‌هایی در مورد حریم خصوصی داده‌ها، بهترین شیوه‌های امنیتی و اهمیت حفاظت از اطلاعات حساس ارائه دهید. افزایش آگاهی می‌تواند به جلوگیری از نشت‌های غیرعمدی داده‌ها کمک کند و فرهنگ امنیتی را تقویت کند.
- **پیاده‌سازی سامانه‌های شناسایی ناهنجاری:**
  - از شناسایی ناهنجاری مبتنی بر یادگیری ماشین برای شناسایی الگوهای غیرمعمول در پرسش‌ها یا دسترسی به مدل استفاده کنید. این می‌تواند به شناسایی سریع حملات استنتاج ویژگی کمک کند و هشدارهایی برای بررسی ایجاد کند.
- **ایجاد یک طرح پاسخ:**
  - یک طرح پاسخ به حوادث به‌ویژه برای نقض‌های داده و حملات استنتاج توسعه و نگهداری کنید. داشتن یک طرح واضح به اقدام سریع برای کاهش آسیب و بازیابی از حمله اجازه می‌دهد.

۳-۶ مسمومیت مدل هوش مصنوعی

مسمومیت مدل[1] هوش مصنوعی به نوعی حمله اشاره دارد که در آن بازیگران مخرب داده‌های آموزشی یک مدل هوش مصنوعی را دستکاری می‌کنند تا عملکرد، یکپارچگی یا قابلیت اطمینان آن را به خطر اندازند. این کار با معرفی داده‌های خراب یا گمراه‌کننده به مجموعه داده‌ای که مدل از آن یاد می‌گیرد، انجام می‌شود و در نتیجه بر رفتار مدل به شیوه‌های ناخواسته تأثیر می‌گذارد.

مسمومیت مدل هوش مصنوعی تهدیدی مهم برای امنیت و اثربخشی سامانه‌های هوش مصنوعی است. درک سازوکارهای آن و توسعه دفاع‌های قوی در برابر چنین حملاتی برای حفظ یکپارچگی

1. Model Poisioning

برنامه‌های هوش مصنوعی حیاتی است.

۳-۶-۱ ویژگی‌های کلیدی

- **دستکاری[1] داده‌ها:**
  - مهاجمان داده‌های مخرب یا گمراه‌کننده را به مجموعه داده آموزشی معرفی می‌کنند تا بر فرآیند یادگیری مدل تأثیر بگذارند.
- **پنهان بودن:**
  - داده‌های مسموم‌شده می‌توانند بسیار شبیه به داده‌های مشروع باشند، که شناسایی آن‌ها را در طول فرآیندهای اعتبارسنجی داده استاندارد دشوار می‌کند.
- **نتایج هدفمند:**
  - حمله می‌تواند به رفتارهای خاصی هدف‌گذاری شود، مانند ایجاد اشتباه در طبقه‌بندی ورودی‌های خاص یا تولید پیش‌بینی‌های جانبدارانه.
- **کاهش عملکرد:**
  - معرفی داده‌های مسموم‌شده می‌تواند منجر به کاهش دقت، افزایش نرخ خطا و کلی کاهش عملکرد مدل شود.
- **تأثیرات بلندمدت:**
  - اثرات مسمومیت می‌تواند پایدار باشد و مدل را حتی پس از حمله تحت تأثیر قرار دهد، اگر به یادگیری از داده‌های آلوده ادامه دهد.
- **آسیب‌پذیری در یادگیری فدرال[2] :**
  - مدل‌های آموزش‌دیده در محیط‌های یادگیری فدرال آسیب‌پذیر هستند، زیرا چندین مشارکت‌کننده ممکن است داده‌های مسموم‌شده را بدون شناسایی وارد کنند.
- **هزینه پایین حمله:**
  - اجرای یک حمله مسمومیت معمولاً به منابع کمتری نسبت به سایر نوع حملات نیاز دارد، که آن را برای طیف وسیع‌تری از مهاجمان قابل دسترس می‌کند.
- **بردارهای حمله متنوع:**
  - مسمومیت می‌تواند در مراحل مختلفی از جمله جمع‌آوری داده، پیش‌پردازش یا در طول

1. Manipulation
2. Federated Learning

مرحله آموزش مدل رخ دهد.

۳-۶-۲ دارایی‌های هدف

- در حملات مسمومیت مدل هوش مصنوعی، دارایی‌های مختلفی می‌توانند هدف قرار گیرند تا اهداف مخرب تحقق یابند. در اینجا دارایی‌های اصلی که مهاجمان معمولاً بر روی آن‌ها تمرکز می‌کنند، آورده شده است:

- **داده‌های آموزشی:**
  - مجموعه داده‌ای که برای آموزش مدل استفاده می‌شود، هدف اصلی است. مهاجمان این داده‌ها را دستکاری می‌کنند تا نمونه‌های مسموم‌شده‌ای معرفی کنند که یادگیری مدل را منحرف می‌کند.
- **مدل‌های یادگیری ماشین:**
  - خود مدل‌ها می‌توانند به‌طور مستقیم هدف قرار گیرند، به‌ویژه در سناریوهایی که شامل محیط‌های یادگیری مشترک یا فدرال هستند، جایی که به‌روزرسانی‌ها می‌توانند مسموم شوند.
- **وزن‌ها و پارامترهای مدل:**
  - مهاجمان ممکن است هدفشان خراب کردن وزن‌ها یا پارامترهای یک مدل آموزش‌دیده باشد که منجر به کاهش عملکرد یا نتایج جانبدارانه می‌شود.
- **داده‌های ورودی:**
  - با مسموم کردن مجموعه داده‌ای که در طول استنتاج استفاده می‌شود، داده‌های ورودی آینده نیز می‌توانند هدف قرار گیرند که می‌تواند به پیش‌بینی‌ها یا طبقه‌بندی‌های نادرست منجر شود.
- **فضای ویژگی:**
  - با دستکاری ویژگی‌های خاص داده‌های آموزشی، مهاجمان می‌توانند الگوهای گمراه‌کننده‌ای ایجاد کنند که مدل یاد می‌گیرد آن‌ها را شناسایی کند و بر پیش‌بینی‌ها و تصمیمات آن تأثیر می‌گذارد.
- **APIها و رابط‌های مدل:**
  - مهاجمان ممکن است از آسیب‌پذیری‌ها در APIها یا رابط‌هایی که از طریق آن‌ها به مدل‌ها دسترسی پیدا می‌شود، سوء استفاده کنند و به این ترتیب اجازه دهند داده‌ها یا به‌روزرسانی‌های مسموم‌شده را تزریق کنند.

۳-۶-۳ آسیب‌پذیری‌ها

- حملات مسمومیت مدل هوش مصنوعی از آسیب‌پذیری‌های خاصی در سامانه‌های یادگیری ماشین بهره‌برداری می‌کنند. در اینجا آسیب‌پذیری‌های کلیدی که می‌توانند هدف قرار گیرند، آورده شده است:

- **وابستگی به داده‌ها:**
  - مدل‌های یادگیری ماشین به شدت به کیفیت و یکپارچگی داده‌های آموزشی خود وابسته هستند. حملات مسمومیت می‌توانند از این وابستگی بهره‌برداری کنند و داده‌های مخرب را معرفی کنند که مدل را منحرف می‌کند.
- **عدم اعتبارسنجی داده:**
  - بسیاری از سامانه‌ها رویه‌های اعتبارسنجی داده قوی را پیاده‌سازی نمی‌کنند، که این امر اجازه می‌دهد مهاجمان به‌راحتی نمونه‌های مسموم‌شده را بدون شناسایی تزریق کنند.
- **پیچیدگی مدل:**
  - مدل‌های پیچیده، مانند شبکه‌های عصبی عمیق، می‌توانند به مسمومیت آسیب‌پذیرتر باشند زیرا ممکن است در مواجهه با داده‌های آموزشی جانبدارانه یا دستکاری‌شده به‌خوبی تعمیم نیابند.
- **ضعف‌های یادگیری فدرال:**
  - در تنظیمات یادگیری فدرال، به‌روزرسانی‌های غیرمتمرکز از چندین شرکت‌کننده می‌توانند مسموم شوند. تجمیع این به‌روزرسانی‌ها بدون اعتبارسنجی مناسب می‌تواند منجر به یک مدل عمومی به خطر افتاده شود.
- **عدم شناسایی ناهنجاری[1] مؤثر:**
  - عدم وجود سازوکارهای شناسایی ناهنجاری مؤثر به مهاجمان اجازه می‌دهد داده‌های مسموم‌شده‌ای را معرفی کنند که در طول آموزش هشدارهای لازم را ایجاد نمی‌کند.
- **اقدامات امنیتی ناکافی:**
  - بسیاری از سامانه‌های هوش مصنوعی پروتکل‌های امنیتی قوی برای دسترسی به داده‌ها و به‌روزرسانی مدل‌ها ندارند و این امر آن‌ها را در برابر تغییرات غیرمجاز آسیب‌پذیر می‌کند.

1. Anomaly Detection

- **بیش‌برازش[1] به داده‌های آموزشی:**
  - مدل‌هایی که به داده‌های آموزشی خود بیش‌برازش می‌کنند ممکن است به مسمومیت آسیب‌پذیرتر باشند، زیرا الگوهایی را یاد می‌گیرند که تحت تأثیر نمونه‌های مسموم‌شده قرار دارند.
- **عوامل انسانی:**
  - توسعه‌دهندگان و دانشمندان داده ممکن است به‌طور ناخواسته با مواردی نظیر تکیه بر منابع غیرقابل اعتماد برای داده‌ها یا عدم ارزیابی دقیق کیفیت داده‌ها آسیب‌پذیری‌هایی را معرفی کنند،
- **شفافیت محدود:**
  - ماهیت غیرشفاف مدل‌های هوش مصنوعی و فرآیندهای آموزشی آن‌ها می‌تواند شناسایی داده‌های مسموم یا تأثیرات مخرب را دشوار کند.

۳-۶-۴ سناریو تهدید

در یک حمله مسمومیت مدل ، هدف مهاجم این است که عملکرد مدل را منحرف کند، یا به‌گونه‌ای که عملکرد ضعیف داشته باشد یا درب‌های پشتی[2] یا سوگیری‌های خاصی را جاسازی کند که مهاجم بتواند در زمان استنتاج از آن‌ها بهره‌برداری کند.

- **شناسایی و دسترسی:**
  - مهاجم اطلاعاتی درباره فرآیند آموزش مدل جمع‌آوری می‌کند، از جمله منابع داده، پروتکل‌های آموزشی، تنظیمات همکاری (مانند یادگیری فدرال) و نحوه اعتبارسنجی و ادغام داده‌ها.
- **تزریق یا دستکاری داده:**
  - در یک سناریوی آموزش متمرکز، مهاجم داده‌های آلوده را به مجموعه داده‌های آموزشی معرفی می‌کند. در یادگیری فدرال، ممکن است تلاش کند تا دستگاه‌های کلاینت را مختل کند تا گرادیان‌های دستکاری‌شده‌ای ارسال کنند.
- **شناسایی تریگرها و شرایط:**
  - مهاجم تریگرهای خاصی را جاسازی می‌کند، مانند الگوهای ورودی خاص که در زمان استنتاج، باعث می‌شوند مدل به شیوه‌ای رفتار کند که مهاجم قصد دارد، مانند طبقه‌بندی نادرست.

1. Overfitting
2. Backdoors

- **عبور از اعتبارسنجی و بررسی‌ها:**
  - مهاجم اطمینان حاصل می‌کند که داده‌ها یا ورودی‌های مخرب او بدون شناسایی از موانع اعتبارسنجی عبور می‌کنند، معمولاً با طراحی ورودی‌هایی که به‌طور آماری مشابه داده‌های مشروع به نظر می‌رسند.
- **تأثیرگذاری بر فرآیند آموزش:**
  - با تزریق هوشمندانه داده‌ها یا گرادیان‌های تعصب‌آمیز، مهاجم به‌روزرسانی وزن‌های مدل را در طول تکرارهای آموزشی تحت تأثیر قرار می‌دهد و سوگیری‌ها یا مرزهای تصمیم‌گیری را تغییر می‌دهد.
- **نظارت و تنظیم:**
  - بسته به پیچیدگی حمله، مهاجم ممکن است به‌طور تکراری خروجی‌های مدل (به‌طور مستقیم یا از طریق آزمایش‌های استنتاج) را نظارت کرده و رویکرد خود را تنظیم کند تا تأثیر نهفته را حفظ یا تقویت کند.

۳-۶-۵ نشانه‌های احتمال نفوذ

در این بخش برخی از نشانه‌های احتمال نفوذ که می‌تواند نشانگر مسمومیت مدل هوش مصنوعی باشد بیان م‌گردد.

- **الگوهای غیرعادی داده:**
  - تغییرات ناگهانی در توزیع داده‌های آموزشی، مانند افزایش غیرمنتظره در برخی کلاس‌ها یا ویژگی‌ها، می‌تواند نشانه‌ای از وجود نمونه‌های مسموم‌شده باشد.
- **کاهش عملکرد مدل:**
  - کاهش قابل توجه در دقت، صحت و یادآوری مدل، به‌ویژه در وظایفی که قبلاً به‌خوبی انجام‌شده، می‌تواند نشان‌دهنده این باشد که مدل به خطر افتاده است.
- **رفتار پیش‌بینی غیرعادی:**
  - مدل شروع به ارائه پیش‌بینی‌هایی می‌کند که با داده‌های تاریخی یا دانش حوزه سازگار نیستند، که این موضوع به احتمال دستکاری داده‌های آموزشی اشاره دارد.
- **برچسب‌گذاری نامتناقض:**
  - ناهماهنگی‌های مکرر در برچسب‌گذاری، مانند افزایش ناگهانی در موارد نادرست طبقه‌بندی‌شده، می‌تواند نشان‌دهنده وجود نمونه‌های تغییر برچسب‌خورده باشد.

- **تغییرات وزن‌های مدل:**
  - تغییرات غیرقابل توضیح در وزن‌ها یا پارامترهای مدل، به‌ویژه پس از به‌روزرسانی‌ها، ممکن است نشان‌دهنده تغییرات غیرمجاز یا مسمومیت باشد.
- **افزایش تأخیر در پاسخ‌های مدل:**
  - افزایش ناگهانی در زمان لازم برای مدل جهت انجام پیش‌بینی‌ها می‌تواند نشان‌دهنده پردازش داده‌های مسموم‌شده یا مشکل در پردازش ورودی‌های خراب باشد.
- **تماس‌های غیرمنتظره API:**
  - نظارت بر الگوهای غیرعادی در درخواست‌ها یا به‌روزرسانی‌های API می‌تواند به شناسایی تلاش‌های غیرمجاز برای تغییر مدل یا تزریق داده‌های مسموم‌شده کمک کند.
- **بازخورد و شکایات کاربران:**
  - افزایش گزارش‌های کاربران درباره پیش‌بینی‌های نادرست یا شکست‌های سامانه می‌تواند به‌عنوان یک هشدار اولیه از مشکلات زیرساختی مرتبط با مسمومیت مدل عمل کند.
- **نقض یکپارچگی منبع داده:**
  - تغییرات در یکپارچگی یا منبع داده‌ها، مانند معرفی مجموعه‌های داده تأییدنشده، می‌تواند با احتمال بالای نشان‌دهنده مسمومیت باشد.
- **تغییرات رفتاری در کاربران یا سامانه‌ها:**
  - تغییرات غیرقابل توضیح در رفتار کاربران یا محیط عملیاتی سامانه می‌تواند نشان‌دهنده نفوذ خارجی یا بهره‌برداری داخلی باشد.

۳-۶-۶ تأثیرات

پیامدهای مسمومیت مدل هوش مصنوعی چندوجهی هستند و می‌توانند بر سازمان‌ها، کاربران و سامانه‌های اجتماعی گسترده‌تر تأثیر بگذارند.

- **کاهش عملکرد مدل:**
  - مدل‌های مسموم ممکن است دقت، صحت و یادآوری کمتری را نشان دهند که منجر به پیش‌بینی‌ها و تصمیمات نادرست می‌شود. این می‌تواند به قابلیت اطمینان سامانه‌های هوش مصنوعی آسیب برساند.
- **خسارات مالی:**
  - کسب‌وکارها ممکن است به دلیل کاهش کارایی عملیاتی، افزایش نرخ خطا یا نیاز به آموزش

مجدد مدل‌ها برای بازیابی از مسمومیت، هزینه‌های قابل توجهی را متحمل شوند.

- **آسیب به شهرت:**
  - سازمان‌های تحت تأثیر مسمومیت مدل ممکن است متحمل آسیب‌های شهرت شوند که منجر به از دست دادن اعتماد مشتری و سهم بازار می‌شود.
- **آسیب‌پذیری‌های امنیتی:**
  - سامانه‌های هوش مصنوعی به خطر افتاده می‌توانند خطرات امنیتی جدیدی را معرفی کنند، به‌ویژه در برنامه‌های بحرانی مانند وسایل نقلیه خودران، بهداشت و درمان یا امنیت سایبری، که ممکن است جان‌ها را به خطر اندازد.
- **پیامدهای قانونی و نظارتی:**
  - اگر سامانه‌های هوش مصنوعی به خطر افتاده باعث آسیب، نقض حریم خصوصی داده‌ها یا تخلف از استانداردهای صنعتی شوند، سازمان‌ها ممکن است با مسئولیت‌های قانونی یا نظارت‌های مقرراتی مواجه شوند.
- **جانبداری و تبعیض:**
  - حملات مسمومیت می‌توانند جانبداری‌ها را در مدل‌های هوش مصنوعی ایجاد یا تشدید کنند که منجر به رفتار ناعادلانه نسبت به افراد یا گروه‌ها، به‌ویژه در حوزه‌های حساسی مانند استخدام یا اجرای قانون می‌شود.
- **اختلال در عملیات:**
  - سازمان‌ها ممکن است به دلیل خرابی‌های مدل در ارائه خدمات دچار اختلال شوند که نیاز به مداخلات اضطراری و تخصیص منابع را ایجاب می‌کند.
- **تأثیر بر یکپارچگی داده:**
  - مجموعه‌های داده آموزشی مسموم می‌توانند یکپارچگی خط لوله داده را به خطر بیندازند و منجر به چالش‌های بلندمدت در حفظ داده‌های با کیفیت بالا شوند.
- **افزایش هزینه‌های منابع:**
  - سازمان‌ها ممکن است نیاز به سرمایه‌گذاری قابل توجهی در اقدامات امنیتی، نظارت و پاسخ به حوادث برای کاهش خطرات مرتبط با مسمومیت مدل‌های هوش مصنوعی داشته باشند.
- **کاهش اعتماد به فناوری‌های هوش مصنوعی:**
  - وقوع گسترده مسمومیت‌ها می‌تواند به بی‌اعتمادی نسبت به فناوری‌های هوش مصنوعی

منجر شود و پذیرش و نوآوری در بخش‌های مختلف را مختل کند.

۳-۶-۷ راه‌های مقابله

برای حفاظت در برابر حملات مسمومیت مدل هوش مصنوعی، سازمان‌ها می‌توانند مجموعه‌ای از تدابیر پیشگیری و مقابله را پیاده‌سازی کنند. پیاده‌سازی این راهبردهای پیشگیری می‌تواند به‌طور قابل توجهی خطر حملات مسمومیت مدل هوش مصنوعی را کاهش دهد و استحکام کلی سامانه‌های هوش مصنوعی را افزایش دهد. رویکردی پیشگیرانه در امنیت و یکپارچگی داده برای حفظ اعتماد و قابلیت اطمینان در فناوری‌های هوش مصنوعی ضروری است

- **اعتبارسنجی داده قوی:**
  - پیاده‌سازی رویه‌های اعتبارسنجی سخت‌گیرانه برای داده‌های آموزشی به منظور اطمینان از یکپارچگی و اصالت آن. این مورد شامل بررسی ناهنجاری‌ها، ناهماهنگی‌ها و الگوهای غیرمنتظره است.
- **سامانه‌های شناسایی ناهنجاری:**
  - استقرار سازوکارهای شناسایی ناهنجاری مبتنی بر یادگیری ماشین برای شناسایی الگوهای غیرعادی در داده‌های آموزشی یا عملکرد مدل، که ممکن است نشان‌دهنده تلاش‌های مسمومیت باشد.
- **تکثیر داده:**
  - استفاده از تکنیک‌های تکثیر داده برای ایجاد مجموعه‌های آموزشی متنوع. این موضوع می‌تواند به کاهش تأثیر هر نمونه مسموم کمک کند.
- **بررسی‌های منظم مدل:**
  - انجام بررسی‌های منظم بر روی مدل‌های هوش مصنوعی به منظور ارزیابی عملکرد آن‌ها و شناسایی هرگونه نشانه‌ای از آسیب یا کاهش به دلیل داده‌های مسموم.
- **امنیت یادگیری فدرال:**
  - پیاده‌سازی تکنیک‌های تجمیع امن در تنظیمات یادگیری فدرال، به منظور کاهش خطر تأثیر به‌روزرسانی‌های مسموم بر مدل عمومی.
- **روش‌های مجموعه‌ای[1]:**
  - استفاده از تکنیک‌های یادگیری مجموعه‌ای که مدل‌های متعددی را ترکیب می‌کند، که باعث

1. Ensemble

می‌شود تأثیر یک مجموعه داده مسموم به‌طور قابل توجهی بر عملکرد کلی سخت‌تر شود.

- **تکنیک‌های آموزشی قوی:**
  - استفاده از روش‌های آموزشی قوی مانند آموزش متخاصم یا استفاده از توابع ضرر قوی. که نسبت به داده‌های پرت و مسموم حساسیت کمتری دارند.
- **کنترل دسترسی و نظارت:**
  - پیاده‌سازی کنترل‌های دسترسی سخت‌گیرانه برای محدود کردن اینکه چه کسی می‌تواند داده‌های آموزشی یا پارامترهای مدل را تغییر دهد و نظارت منظم بر گزارش‌های دسترسی برای شناسایی فعالیت‌های مشکوک.
- **آموزش و آگاهی کاربران:**
  - آموزش کاربران و ذی‌نفعان درباره خطرات مسمومیت مدل و اهمیت کیفیت داده، و تشویق به دقت در روش‌های مدیریت داده.
- **برنامه‌ریزی پاسخ به حوادث:**
  - توسعه و نگهداری یک برنامه پاسخ به حادثه به‌طور خاص برای مسمومیت مدل هوش مصنوعی که مراحل شناسایی، مهار و بازیابی را مشخص می‌کند.
- **نسخه‌گذاری مدل:**
  - پیاده‌سازی کنترل نسخه برای مدل‌ها به منظور تسهیل بازگشت به نسخه‌های قبلی و غیرمسموم در صورت شناسایی حمله مسمومیت.
- **تنوع در منابع داده:**
  - استفاده از منابع مختلف و متنوع برای داده‌های آموزشی به منظور کاهش خطر از یک منبع مسموم که ممکن است کل مجموعه داده را به خطر بیندازد.

۳-۷ مسمومیت با برچسب تمیز

مسمومیت با برچسب تمیز[1] یک نوع حمله پیشرفته به مدل‌های یادگیری ماشین است که در آن یک مهاجم داده‌های مخرب را به مجموعه داده آموزشی تزریق می‌کند بدون اینکه برچسب‌های قابل مشاهده را تغییر دهد. برخلاف حملات مسمومیت سنتی که برچسب‌های نادرست یا داده‌های به‌وضوح خراب را معرفی می‌کنند، مسمومیت با برچسب تمیز از داده‌های ظاهراً مشروع استفاده می‌کند که می‌تواند مدل را در طول آموزش فریب دهد و اغلب بدون شناسایی فوری انجام می‌شود.

1. Clean Label Poisening

۳-۷-۱ ویژگی‌های کلیدی

- **برچسب‌های مشروع:**
  - داده‌های مسموم برچسب‌های صحیح را حفظ می‌کنند، که شناسایی ناهنجاری‌ها را برای سامانه‌های تشخیص دشوار می‌سازد. این به مهاجم اجازه می‌دهد تا فرآیند یادگیری مدل را دستکاری کند در حالی که ظاهر یکپارچگی داده را حفظ می‌کند.
- **دستکاری ظریف:**
  - مهاجمان معمولاً تنها جنبه‌های کوچک داده را تغییر می‌دهند، مانند مقادیر پیکسل در تصاویر یا تغییرات جزئی در متن. تغییرات ظریف می‌توانند مدل را گمراه کنند بدون اینکه مشکوک به نظر برسند.
- **تأثیر هدفمند:**
  - مسمومیت با برچسب تمیز معمولاً به‌دنبال گمراه کردن مدل برای ایجاد خطاها یا سوگیری‌های خاص است، مانند طبقه‌بندی نادرست کلاس‌های خاص. این رویکرد هدفمند می‌تواند عواقب قابل توجهی برای تصمیم‌گیری در برنامه‌های حیاتی داشته باشد.
- **افزایش پیچیدگی برای شناسایی:**
  - به‌دلیل اینکه برچسب‌ها تغییر نمی‌کنند، تکنیک‌های معمول تشخیص ناهنجاری ممکن است نتوانند داده‌های مسموم را شناسایی کنند. این موضوع شناسایی مجموعه‌های داده آسیب‌دیده را پیچیده می‌کند و نیاز به راهبردهای شناسایی پیشرفته‌تری دارد.
- **استفاده از آسیب‌پذیری‌های مدل:**
  - مهاجمان از ضعف‌های شناخته‌شده در مدل‌های یادگیری ماشین، مانند بیش‌برازش یا حساسیت به نویز، بهره‌برداری می‌کنند. این بهره‌برداری می‌تواند منجر به کاهش شدید عملکرد یا پیش‌بینی‌های جانبدارانه شود.
- **پتانسیل برای حملات با تأثیر بالا:**
  - مسمومیت با برچسب تمیز می‌تواند منجر به اثرات منفی قابل توجهی، به‌ویژه در حوزه‌های حساس مانند مالی، بهداشت و سامانه‌های خودران شود. توانایی دستکاری نتایج در حالی که به‌نظر مشروع می‌رسند می‌تواند عواقب وخیمی به همراه داشته باشد.
- **استفاده از مدل‌های مولد[1]:**
  - مهاجمان ممکن است از مدل‌های مولد برای ایجاد داده‌های مسموم استفاده کنند که به‌طور

1. Generative Model

یکپارچه در مجموعه داده‌های موجود جا بگیرد. این موضوع احتمال این که داده‌های مسموم در حین آموزش نادیده گرفته شوند را افزایش می‌دهد.

- **سختی در انتساب:**
  - تعیین منبع مسمومیت با برچسب تمیز به‌دلیل عدم وجود نشانه‌های واضح از دستکاری، می‌تواند دشوار باشد. این موضوع پاسخ به حملات را پیچیده می‌کند و پیاده‌سازی تدابیر مؤثر را دشوارتر می‌سازد.
- **دستکاری هدفمند کلاس‌ها:**
  - مهاجمان می‌توانند به‌طور انتخابی داده‌های مسموم را وارد کنند تا بر کلاس‌های خاص تأثیر بگذارند، که منجر به نتایج طبقه‌بندی جانبدارانه می‌شود. این می‌تواند پیش‌بینی‌های مدل را به نفع یا علیه گروه‌های خاصی منحرف کند و نگرانی‌های اخلاقی و انصاف را به همراه داشته باشد.
- **اثرهای بلندمدت:**
  - اثرات مسمومیت با برچسب تمیز ممکن است بلافاصله مشهود نباشد، که منجر به کاهش تدریجی عملکرد مدل می‌شود. این تأثیر تأخیری می‌تواند شناسایی و تلاش‌های جبران خسارت را مختل کند.

۳-۷-۲ دارایی‌های هدف

در حملات مسمومیت با برچسب تمیز، مهاجمان به‌دنبال دستکاری دارایی‌های خاصی در سامانه‌های یادگیری ماشین هستند بدون اینکه برچسب‌های قابل مشاهده داده را تغییر دهند. در اینجا اهداف اصلی دارایی‌های مسمومیت با برچسب تمیز آورده شده است:

- **مجموعه‌های داده آموزشی:**
  - دارایی اصلی هدف‌گذاری‌شده توسط مسمومیت با برچسب تمیز، جایی است که مهاجمان نمونه‌های مخرب را معرفی می‌کنند. مجموعه‌های داده آسیب‌دیده می‌توانند منجر به آموزش نادرست مدل شوند و پیش‌بینی‌های غیرقابل اعتمادی را به همراه داشته باشند.
- **مدل‌های یادگیری ماشین:**
  - مدل‌های موجود که به یکپارچگی داده‌های آموزشی وابسته هستند، می‌توانند به‌طور غیرمستقیم از طریق ورودی‌های مسموم هدف قرار گیرند. این می‌تواند عملکرد مدل را کاهش دهد، سوگیری‌ها را معرفی کند یا به طبقه‌بندی‌های نادرست منجر شود.

- **اهمیت ویژگی:**
  - مهاجمان ممکن است به‌دنبال دستکاری در درک مدل از اینکه کدام ویژگی‌ها مهم هستند، با معرفی نمونه‌های با برچسب تمیز که اهمیت ویژگی را تحریف می‌کنند، باشند. این می‌تواند تصمیمات مدل را منحرف کند و انجام اهداف خاص حمله را آسان‌تر کند.
- **معیارهای ارزیابی:**
  - معیارهای استفاده‌شده برای ارزیابی عملکرد مدل می‌توانند هدف قرار گیرند با معرفی داده‌هایی که به‌طور مصنوعی این معیارها را افزایش می‌دهند. نتایج ارزیابی گمراه‌کننده می‌توانند بر فرآیندهای تصمیم‌گیری تأثیر بگذارند و منجر به اعتماد نادرست به عملکرد مدل شوند.
- **ورودی‌های کاربر:**
  - در سامانه‌هایی که بر اساس بازخورد کاربر سازگار می‌شوند، مهاجمان ممکن است ورودی‌های مسموم‌شده‌ای را معرفی کنند که به‌نظر مشروع می‌رسند. این می‌تواند منجر به خطاهای تجمعی شود زیرا مدل از داده‌های آلوده کاربر یاد می‌گیرد.
- **خطوط لوله داده[1]:**
  - فرآیندهایی که از طریق آن‌ها داده‌ها به آموزش و ارزیابی وارد می‌شوند می‌توانند هدف قرار گیرند تا داده‌های مسموم در مراحل مختلف معرفی شوند. خطوط لوله داده آسیب‌دیده می‌توانند اثرات مسمومیت را در طول چرخه عمر یادگیری ماشین تداوم بخشند.
- **محیطهای استقرار:**
  - محیطهایی که مدل‌ها در آن‌ها مستقر می‌شوند می‌توانند هدف قرار گیرند تا اطمینان حاصل شود که داده‌های مسموم به‌طور مداوم به سامانه وارد می‌شوند. این می‌تواند منجر به دستکاری مداوم پیش‌بینی‌های مدل در برنامه‌های زمان واقعی شود.
- **سازوکارهای بازخورد:**
  - سامانه‌هایی که از تعاملات و بازخوردهای کاربران یاد می‌گیرند می‌توانند با معرفی بازخوردهای به‌ظاهر دقیق که در واقع مسموم هستند، مورد سوءاستفاده قرار گیرند. این امر فرآیند یادگیری را تضعیف می‌کند و می‌تواند باعث کاهش کیفیت مدل در طولانی‌مدت شود.
- **رعایت مقررات:**
  - مسمومیت با برچسب تمیز می‌تواند معیارهای انطباق و فرآیندهای گزارش‌گیری که به

1. Data Pipelines

خروجی‌های مدل وابسته است، هدف قرار دهد که ممکن است مقامات را گمراه کند. این می‌تواند به پیامدهای قانونی و آسیب به اعتبار سازمان منجر شود.

- **شهرت و اعتماد:**
  - با به خطر انداختن یکپارچگی مدل‌ها و داده‌ها، مهاجمان می‌توانند به شهرت یک سازمان آسیب بزنند. از دست دادن اعتماد از سوی کاربران و ذینفعان می‌تواند تأثیرات ماندگاری بر قابلیت بقا کسب و کار و روابط با مشتریان داشته باشد.

۳-۷-۳ آسیب‌پذیری‌ها

مسمومیت با برچسب تمیز از آسیب‌پذیری‌های خاصی در سامانه‌های یادگیری ماشین بهره‌برداری می‌کند که آن را به نوعی حمله دسیسه آمیز تبدیل می‌کند. در اینجا آسیب‌پذیری‌های کلیدی که تسهیل‌کننده مسمومیت با برچسب تمیز هستند، آورده شده است:

- **بیش‌برازش مدل:**
  - مدل‌هایی که به‌طور دقیق و بیش از حد بر روی داده‌های آموزشی خود آموزش دیده‌اند ممکن است به تغییرات ظریف در داده‌های ورودی حساس شوند. این حساسیت به مهاجمان اجازه می‌دهد که تغییرات کوچک و به‌ظاهر بی‌ضرر را معرفی کنند که می‌تواند پیش‌بینی‌های مدل را به‌طور قابل توجهی تغییر دهد.
- **ضعف در استحکام[1]:**
  - بسیاری از الگوریتم‌های یادگیری ماشین به‌طور مؤثر برای مقابله با ورودی‌های متخاصم یا داده‌های نویزدار طراحی نشده‌اند. این آسیب‌پذیری ذاتی، مدل‌ها را به‌طور خاصی در برابر دستکاری از طریق مسمومیت با برچسب تمیز آسیب‌پذیر می‌کند.
- **ضعف در کیفیت داده:**
  - فرآیندهای ناکافی اعتبارسنجی و پاک‌سازی داده می‌توانند اجازه دهند که داده‌های مسموم به مجموعه‌های آموزشی وارد شوند. حکمرانی ضعیف داده، خطر شناسایی نشدن مجموعه‌های داده آسیب‌دیده را افزایش می‌دهد.
- **عدم تشخیص ناهنجاری کافی:**
  - تکنیک‌های استاندارد تشخیص ناهنجاری ممکن است به‌دلیل برچسب‌های مشروع داده‌های مسموم نتوانند حملات با برچسب تمیز را شناسایی کنند. این موضوع به مهاجمان اجازه می‌دهد که داده‌های مضر را بدون ایجاد هشدار در حین فرآیند آموزش معرفی کنند.

1. Robustness

- **حلقه‌های بازخورد:**
  - سامانه‌هایی که از ورودی یا بازخورد کاربران یاد می‌گیرند می‌توانند مورد سوءاستفاده قرار گیرند اگر مهاجمان پاسخ‌های مسمومشده‌ای را معرفی کنند که به‌نظر معتبر می‌رسند. این می‌تواند منجر به کاهش تجمعی عملکرد مدل در طول زمان شود.
- **حساسیت به ویژگی:**
  - مدل‌هایی که به‌طور جدی به ویژگی‌های خاصی وابسته هستند در صورتی که داده‌های مسموم، اهمیت درک‌شده آن ویژگی‌ها را تغییر دهند می‌توانند به‌راحتی دستکاری شوند. مهاجمان می‌توانند از این حساسیت برای منحرف کردن تصمیمات مدل به نفع خود بهره‌برداری کنند.
- **ضعف‌های آماری:**
  - بسیاری از مدل‌ها به فرضیات آماری وابسته‌اند که می‌تواند با معرفی داده‌های مسموم نقض شود. اگر فرضیات بنیادی مدل به خطر بیفتد، پیش‌بینی‌های آن می‌تواند غیرقابل اعتماد شود.
- **عدم توضیح کافی:**
  - اغلب مدل‌های یادگیری ماشین، به‌ویژه مدل‌های یادگیری عمیق، از عدم قابلیت تفسیر و قابلیت توضیح رنج می‌برند که شناسایی دلایل تصمیم‌گیری را دشوار می‌کند. این ابهام می‌تواند مانع شناسایی ورودی‌های دستکاری‌شده شده و منجر به آسیب‌پذیری‌های پایدار شود.
- **آزمایش و اعتبارسنجی ناکافی:**
  - پروتکل‌های آزمایش و اعتبارسنجی ناکافی می‌توانند سامانه‌ها را در برابر مسمومیت با برچسب تمیز آسیب‌پذیر کنند. این بی‌توجهی می‌تواند منجر به استقرار مدل‌ها بدون درک واضح از ضعف‌های آن‌ها شود.
- **وابستگی به داده‌های تاریخی:**
  - مدل‌هایی که به‌طور جدی به داده‌های تاریخی وابسته هستند می‌توانند قدیمی شوند و به انواع جدیدی از مسمومیت داده آسیب‌پذیر شوند که در طول آموزش وجود نداشته‌اند. این امر می‌تواند منجر به دستکاری مدل‌ها توسط نمونه‌های متخاصم جدیدی شود که از ویژگی‌های مجموعه داده اصلی منحرف هستند.

۳-۷-۴ سناریو تهدید

در یک حمله مسمومیت با برچسب تمیز، مهاجم هدف دارد تا تغییرات ظریفی را در داده‌های آموزشی

جاسازی کند که منجر به شکست‌ها یا سوگیری‌ها خاصی در رفتار مدل شود. این نقاط داده‌های آلوده باید به‌گونه‌ای به نظر برسند که بی‌خطر و غیرقابل تشخیص از داده‌های آموزشی عادی باشند تا از شناسایی در طول پیش‌پردازش و اعتبارسنجی جلوگیری شود. این حمله می‌تواند به‌ویژه در شرایطی مؤثر باشد که داده‌ها از منابع جمع‌سپاری‌شده یا منابع خارجی تأییدنشده می‌آیند.

- **تحلیل منبع داده:**
  - مهاجم فرآیندها و منابع بالقوه جمع‌آوری داده را شناسایی می‌کند، به‌ویژه آن‌هایی که به مجموعه آموزشی مدل داده‌های جدید اضافه می‌کنند. هدف پیدا کردن نقاطی است که در آن‌ها بررسی‌های یکپارچگی داده حداقلی یا غیاب دارد.
- **انتخاب هدف:**
  - مهاجم نتایج یا رفتارهای خاصی را که می‌خواهد در مدل ایجاد کند، انتخاب می‌کند. این می‌تواند شامل هدف‌گذاری بر روی کلاس‌های خاصی باشد که هنگام حضور ویژگی‌ها یا الگوهای خاص، به‌طور نادرست طبقه‌بندی شوند.
- **ساخت داده‌های آلوده:**
  - مهاجم به‌دقت نمونه‌های داده‌ای را می‌سازد که به نظر معتبر می‌رسند و به یک کلاس تعلق دارند، اما به‌طور ظریف ویژگی‌هایی را شامل می‌شوند که آن‌ها را به کلاس هدف دیگری متصل می‌کند. این تغییرات باید برای ناظران انسانی یا تکنیک‌های اعتبارسنجی استاندارد غیرقابل تشخیص باشند.
- **تزریق داده:**
  - نمونه‌های ساخته‌شده به مجموعه داده‌های آموزشی اضافه می‌شوند. با توجه به ماهیت ظریف تغییرات، این نمونه‌ها از عملیات پاک‌سازی داده‌های عادی عبور می‌کنند بدون اینکه مشکوک به نظر برسند.
- **آموزش مدل و تأثیر مسمومیت:**
  - با آموزش مدل بر روی این مجموعه داده، نمونه‌های آلوده بر مرزهای تصمیم‌گیری تأثیر می‌گذارند و منجر به مدلی می‌شوند که تحت شرایط خاص یا الگوهای ورودی تعریف‌شده توسط مهاجم به‌طور نادرست رفتار می‌کند.

۳-۷-۵ نشانه‌های احتمال نفوذ

شناسایی مسمومیت با برچسب تمیز به‌دلیل ظرافت تغییرات ایجادشده در داده‌ها می‌تواند چالش‌برانگیز باشد. با این حال، چندین نشانه احتمال نفوذ وجود دارد که می‌تواند به سازمان‌ها

در شناسایی حملات بالقوه مسمومیت با برچسب تمیز کمک کند. در ادامه نشانه‌های کلیدی برای نظارت آورده شده است:

- **رفتار غیرمنتظره مدل:**
  - انحرافات قابل توجه در پیش‌بینی‌های مدل یا معیارهای عملکرد، به‌ویژه در کلاس‌های خاص. ناهنجاری‌ها در دقت، صحت یا نرخ یادآوری که در ارزیابی‌های قبلی وجود نداشتند.
- **افزایش طبقه‌بندی‌های نادرست:**
  - افزایش نرخ‌های طبقه‌بندی نادرست، به‌خصوص برای دسته‌هایی که قبلاً پایدار بودند. افزایش قابل توجه در مثبت‌های کاذب یا منفی‌های کاذب در خروجی‌های مدل نشانه‌های متصور است.
- **شناسایی انحراف داده:**
  - تغییرات در توزیع داده که با روندها یا به‌روزرسانی‌های مورد انتظار مطابقت ندارد. آزمایش‌های آماری که تغییرات معنی‌داری در داده‌های ورودی نسبت به الگوهای تاریخی نشان می‌دهد نشانه‌ای محتمل برای این مورد است.
- **اهمیت غیرعادی ویژگی:**
  - تغییرات در اهمیتی که به ویژگی‌ها اختصاص داده شده است که با دانش دامنه یا رفتار قبلی مدل ناسازگار است. افزایش یا کاهش ناگهانی در نمرات اهمیت ویژگی برای پیش‌بین‌های کلیدی از مهم‌ترین نشانه‌ها هستند.
- **الگوهای غیرمعمول در داده‌های ورودی:**
  - معرفی نقاط داده‌ای که الگوها یا ویژگی‌های غیرمعمولی در مقایسه با بقیه مجموعه داده نشان می‌دهند. از جمله نشانه‌های محتمل می‌توان به ناهنجاری‌هایی که از طریق تکنیک‌های خوشه‌بندی یا شناسایی ناهنجاری شناسایی می‌شوند اشاره کرد.
- **ناهنجاری‌های حلقه بازخورد:**
  - ناهنجاری‌ها در بازخورد یا تعاملات کاربر، به‌ویژه اگر از رفتار مورد انتظار کاربر منحرف شود. تغییرات ناگهانی در رتبه‌بندی‌های کاربر یا روندهای بازخورد که با تغییرات محصول یا خدمات همخوانی ندارد.
- **تغییرات در رفتار کاربر:**
  - تغییر غیرقابل توضیح در تعاملات کاربر که ممکن است نشان‌دهنده دستکاری از طریق داده‌های مسموم باشد. الگوهای استفاده یا تعامل که از نُرم‌های تاریخی منحرف می‌شوند.

- **اختلال در بررسی‌های یکپارچگی داده:**
  - شکست در بررسی‌های اعتبارسنجی داده که باید مشکلات بالقوه را شناسایی کند و اخطارهایی که توسط ابزارهای نظارت بر کیفیت داده صادر می‌شود و نشان‌دهنده ناهنجاری‌های غیرمنتظره است.
- **افزایش نویز در داده‌های آموزشی:**
  - افزایش سطح نویز در مجموعه داده‌های آموزشی که می‌تواند سیگنال را تضعیف کرده و بر یادگیری مدل تأثیر بگذارد و معیارهای آماری که افزایش تغییرپذیری یا عدم انسجام در کیفیت داده را نشان می‌دهد.
- **مقایسه با معیارهای خارجی:**
  - تفاوت‌ها بین عملکرد مدل و معیارهای خارجی یا مدل‌های مشابه و شکاف‌های قابل توجه در عملکرد در مقایسه با استانداردهای صنعتی یا مدل‌های همتایان.

۳-۷-۶ تأثیرات

- مسمومیت با برچسب تمیز می‌تواند پیامدهای وسیعی در حوزه‌های مختلف داشته باشد. در اینجا برخی از تأثیرات کلیدی مرتبط با حملات مسمومیت با برچسب تمیز آورده شده است:
- **کاهش عملکرد مدل:**
  - معرفی داده‌های مسموم می‌تواند منجر به کاهش دقت، صحت و یادآوری در مدل‌های یادگیری ماشین شود. این کاهش می‌تواند منجر به تصمیم‌گیری‌های ضعیف و خروجی‌های غیرقابل اعتماد شود که بر اعتماد و رضایت کاربر تأثیر می‌گذارد.
- **سوگیری:**
  - مسمومیت با برچسب تمیز می‌تواند پیش‌بینی‌های مدل را منحرف کرده و منجر به نتایج جانبدارانه‌ای شود که به نفع یا ضرر گروه‌های خاص است. این می‌تواند به تبعیض در برنامه‌هایی مانند الگوریتم‌های استخدام، تأیید وام، یا سامانه‌های اجرای قانون منجر شود.
- **زیان مالی:**
  - سازمان‌ها ممکن است با عواقب مالی قابل توجهی روبرو شوند که شامل کاهش درآمد و افزایش هزینه‌های عملیاتی می‌شود چرا که مدل‌های آسیب‌دیده می‌توانند به تصمیمات مالی نادرست منجر شوند، مانند راهبردهای تجاری نادرست یا شکست در شناسایی تقلب.

- **آسیب به شهرت:**
  - شرکت‌ها ممکن است در صورت تولید خروجی‌های غیرقابل اعتماد یا جانبدارانه به دلیل داده‌های مسموم، آسیب‌های شهرتی متحمل شوند. از دست دادن اعتماد مشتری می‌تواند منجر به کاهش سهم بازار، ترک مشتریان و دشواری در جذب مشتریان جدید شود.
- **پیامدهای قانونی و نظارتی:**
  - سازمان‌ها ممکن است با عواقب قانونی و نظارتی روبرو شوند اگر مدل‌های جانبدار یا نادرست در حوزه‌های حیاتی به کار گرفته شوند.این می‌تواند منجر به جریمه‌ها، تحریم‌ها و افزایش نظارت از سوی نهادهای نظارتی شود.
- **اختلال در عملیات:**
  - مسمومیت با برچسب تمیز می‌تواند منجر به اختلالات در سامانه‌های خودکار شود که منجر به انجام اقدامات نادرست بر اساس پیش‌بینی‌های جانبی می‌شود که می‌تواند بر ارائه خدمات، لجستیک و کارایی کلی عملیات تأثیر بگذارد.
- **آسیب به امنیت:**
  - اگر مدل‌ها برای مقاصد امنیتی (مانند شناسایی نفوذ) استفاده شوند، داده‌های مسموم می‌توانند منجر به تهدیدات شناسایی‌نشده شوند. و این آسیب‌پذیری به حملات سایبری را افزایش داده و امنیت سازمان را به خطر می‌اندازد.
- **تأثیر بر تجربه کاربر:**
  - کاربران نهایی ممکن است به‌دلیل پیش‌بینی‌های غیرقابل اعتماد مدل، افت کیفیت خدمات را تجربه کنند. این می‌تواند منجر به نارضایتی کاربر، بازخورد منفی و کاهش تعامل با پلتفرم یا خدمات شود.
- **عقب‌ماندگی در تحقیقات و توسعه:**
  - مسمومیت با برچسب تمیز می‌تواند پیشرفت تلاش‌های تحقیق و توسعه در یادگیری ماشین و هوش مصنوعی را مختل کندو ممکن است منابع به‌منظور رسیدگی به پیامدهای ناشی از مدل‌های آسیب‌دیده صرف شود تا اینکه به پیشرفت نوآوری بپردازند.

۳-۷-۷ راه‌های مقابله

برای مقابله مؤثر با حملات مسمومیت با برچسب تمیز، سازمان‌ها می‌توانند چندین راهبرد کاهش آسیب را پیاده‌سازی کنند. در اینجا رویکردهای کلیدی برای افزایش تاب‌آوری سامانه‌های یادگیری ماشین آورده شده است:

- **اعتبارسنجی داده‌های قوی:**
  - پیاده‌سازی فرآیندهای اعتبارسنجی دقیق برای داده‌های ورودی به‌منظور اطمینان از یکپارچگی و کیفیت آن‌ها و استفاده از تکنیک‌های آماری و شناسایی ناهنجاری‌ها برای شناسایی و فیلتر کردن نقاط داده مشکوک یا خارج از محدوده.
- **ممیزی‌های منظم مدل:**
  - انجام ممیزی‌های مکرر از مدل‌های یادگیری ماشین به‌منظور ارزیابی عملکرد و شناسایی تعصب‌ها یا ناهنجاری‌های بالقوه و مرور منظم خروجی‌های مدل و مقایسه آن‌ها با رفتارهای مورد انتظار برای شناسایی ناهنجاری‌ها به‌موقع.
- **آموزش متخاصم:**
  - گنجاندن نمونه‌های متخاصم در فرآیند آموزش به‌منظور کمک به مدل‌ها برای مقاوم‌تر شدن در برابر دستکاری. با قرار دادن مدل‌ها در معرض ورودی‌های مختل شده متنوع، آن‌ها می‌توانند یاد بگیرند که تلاش‌های مسمومیت با برچسب تمیز را شناسایی و مقاومت کنند.
- **نظارت و شناسایی ناهنجاری:**
  - پیاده‌سازی سامانه‌های نظارت در زمان واقعی برای پیگیری عملکرد مدل و ورودی‌های داده. استفاده از تکنیک‌های شناسایی ناهنجاری مبتنی بر یادگیری ماشین برای علامت‌گذاری الگوهای غیرمعمول که ممکن است نشان‌دهنده مسمومیت باشند.
- **تنوع ورودی داده:**
  - استفاده از مجموعه‌های داده بزرگ و متنوع برای آموزش مدل‌ها، که آن‌ها را کمتر در معرض تأثیر ورودی‌های مسموم قرار می‌دهد. گنجاندن داده‌های متنوع می‌تواند اثر نقاط داده مخرب را تضعیف کند.
- **انتخاب ویژگی قوی:**
  - انتخاب دقیق ویژگی‌هایی که کمتر به دستکاری حساس هستند و ارتباط واضحی با متغیر هدف دارند. این می‌تواند سطح حمله برای مسمومیت با برچسب تمیز را کاهش دهد و وابستگی به ویژگی‌های آسان برای دستکاری را به حداقل برساند.
- **یادگیری ترکیبی:**
  - استفاده از روش‌های ترکیبی که چندین مدل را برای پیش‌بینی ترکیب می‌کنند و در نتیجه وابستگی به هر مدل به‌طور جداگانه را کاهش می‌دهد. این موضوع می‌تواند به کاهش تأثیر داده‌های مسموم کمک کند، زیرا مدل‌های مختلف ممکن است به ورودی‌های مشابه

به‌گونه‌ای متفاوت واکنش نشان دهند.

- **هوش مصنوعی قابل توضیح[1]:**
  - پیاده‌سازی تکنیک‌های هوش مصنوعی قابل توضیح به‌منظور درک بهتر فرآیندهای تصمیم‌گیری مدل. این می‌تواند به شناسایی رفتار غیرمعمول مدل که ممکن است ناشی از مسمومیت با برچسب تمیز باشد، کمک کند و مداخله‌های سریع‌تری را تسهیل کند.
- **سازوکارهای بازخورد:**
  - ایجاد حلقه‌های بازخورد قوی با کاربران برای شناسایی و گزارش رفتار غیرمنتظره مدل یا خروجی‌های نادرست. بینش‌های کاربران می‌توانند اطلاعات ارزشمندی برای شناسایی احتمال آسیب‌پذیری در عملکرد مدل فراهم کنند.
- **همکاری و به اشتراک‌گذاری اطلاعات:**
  - مشارکت در همکاری‌ها با همتایان صنعتی به‌منظور به اشتراک‌گذاری دانش و بهترین شیوه‌ها در مورد تهدیدات مسمومیت با برچسب تمیز. این جمع‌سپاری می‌تواند دفاع‌ها را تقویت کرده و قابلیت‌های شناسایی را در سطح صنعت بهبود بخشد.

۳-۸ مسمومیت هدفمند

حملات مسمومیت هدفمند[2] در زمینه یادگیری ماشین نوع خاصی از حملات مسمومیت داده هستند که به منظور کاهش عملکرد یک مدل در ورودی‌های خاص طراحی شده‌اند یا به مدل القا می‌کنند که خروجی‌های نادرستی برای نمونه‌های هدف خاص تولید کند، در حالی که عملکرد کلی مدل حفظ می‌شود. برخلاف حملات مسمومیت غیرانتخابی که به طور کلی دقت مدل را کاهش می‌دهند، حملات هدفمند بر روی دستکاری رفتار مدل در سناریوهای خاص و انتخاب‌شده به‌طور خصمانه متمرکز هستند.

۳-۸-۱ ویژگی‌های کلیدی

حملات مسمومیت هدفمند با چندین ویژگی متمایز مشخص می‌شوند که آن‌ها را از راهبردهای گسترده‌تر مسمومیت داده متمایز می‌کند. این ویژگی‌های کلیدی شامل دامنه، هدف و اجرای حمله هستند:

- **دقت و خاص بودن:**
  - حمله به گونه‌ای طراحی شده که بر ورودی‌ها یا دسته‌های خاصی از داده‌ها تأثیر بگذارد و نحوه یادگیری مدل برای مدیریت این موارد خاص را دستکاری کند و تنها برخی از نمونه‌ها

1. Explainble AI
2. Targeted Poisoning

یا کلاس‌ها به‌طور نادرست شناسایی یا دستکاری می‌شوند، در حالی که بقیه عملکرد مدل به‌طور کلی تحت تأثیر قرار نمی‌گیرد.

- **زیرکی و پنهان‌کاری:**
  - تغییرات در مجموعه داده‌ها حداقلی و به‌دقت طراحی شده‌اند تا در طول فرآیندهای استاندارد اعتبارسنجی داده و تست مدل نادیده گرفته شوند. هدف حمله دور زدن سازوکارهای شناسایی است به‌طوری‌که معیارهای عملکرد کلی مدل نشان‌دهنده کاهش قابل توجهی نباشد.
- **طبقه‌بندی نادرست عمدی:**
  - حمله‌کننده به‌دنبال ایجاد اشتباهات خاص در شناسایی یا خروجی‌ها تنها برای نمونه‌های هدف است که با اهداف خصمانه‌ای مانند دور زدن تدابیر امنیتی یا تأثیر بر فرآیندهای تصمیم‌گیری هم‌راستا است.
- **تأثیر محدود بر عملکرد کلی:**
  - حمله به‌گونه‌ای طراحی شده که تأثیر کم یا هیچ‌گونه‌ای بر دقت و عملکرد کلی مدل نداشته باشد، که شناسایی آن را از طریق ارزیابی‌های عملکردی سطح بالا دشوار می‌سازد. با عدم کاهش عملکرد کلی سامانه، حمله از جلب شک و بررسی فوری جلوگیری می‌کند.
- **دستکاری پنهانی:**
  - داده‌های آلوده به‌گونه‌ای طراحی شده‌اند که به‌طور یکپارچه با داده‌های معتبر ترکیب شوند، با استفاده از تکنیک‌هایی مانند معرفی نویز ملایم، تغییر جزئی برچسب‌ها یا تغییرات هدفمند در ویژگی‌ها تا شناسایی نشوند.
- **استفاده از آسیب‌پذیری‌های مدل:**
  - حمله ممکن است از نقاط ضعف یا ویژگی‌های خاص معماری مدل که مستعد دستکاری از طریق نمونه‌های به‌دقت طراحی شده هستند، بهره‌برداری کند.

این ویژگی‌ها حملات مسمومیت هدفمند را به طور ویژه‌ای خطرناک می‌سازند، زیرا عملکرد خاصی از مدل یادگیری ماشین را بدون علائم ظاهری تضعیف می‌کنند و چالش‌های قابل توجهی در شناسایی و پیشگیری به وجود می‌آورند.

۳-۸-۲ دارایی‌های هدف

- **داده‌های آموزشی:**
  - این حملات معمولاً بر روی نمونه‌های خاصی در مجموعه داده تمرکز دارند و آن‌ها را تغییر

می‌دهند تا بر درک مدل از نمونه‌های مشابه در حین استنتاج تأثیر بگذارند. نادرست برچسب‌گذاری کردن نقاط داده خاص می‌تواند منجر به یادگیری ارتباطات نادرست توسط مدل شود که پیش‌بینی‌های آن را برای آن نمونه‌های خاص تحت تأثیر قرار می‌دهد.

- **فرآیند آموزش مدل:**
  - با تغییر ویژگی‌های نقاط داده هدف، مهاجمان می‌توانند فرآیند یادگیری ویژگی‌ها را منحرف کنند و باعث شوند مدل پیش‌بینی‌های نادرستی برای آن ویژگی‌ها انجام دهد. مدل‌هایی که به‌طور بیش از حد به الگوهای خاص یا داده‌های دورافتاده وابسته هستند، می‌توانند با معرفی الگوهای مسموم هدفمند، که یادگیری را گمراه می‌کنند، تحت تأثیر قرار گیرند.
- **معماری مدل:**
  - برخی معماری‌ها اگر داده‌های مسموم به‌طور هوشمندانه طراحی شده باشند. ممکن است در برابر یادگیری سوگیری‌ها یا خطاهای خاص آسیب‌پذیرتر باشند. مدل‌هایی که سازوکارهایی برای فیلتر کردن یا کاهش تأثیر داده‌های ناهنجار ندارند، ممکن است بیشتر در برابر حملات هدفمند آسیب‌پذیر شوند.
- **خطوط لوله داده و پیش‌پردازش:**
  - مهاجمان ممکن است فرآیندهایی که داده را تقویت یا تغییر می‌دهند، دستکاری کنند و تغییرات ملایمی را در طول پیش‌پردازش وارد کنند تا بر آموزش مدل تأثیر بگذارند. هدف قرار دادن مراحل پردازش داده که در آن ویژگی‌ها استخراج می‌شوند ممکن است به مهاجمان اجازه دهد تا نویز یا ویژگی‌های گمراه‌کننده را به‌طور انتخابی وارد کنند.
- **پروتکل‌های تست و اعتبارسنجی:**
  - با درک محدودیت‌ها یا نقاط کور در پروتکل‌های تست کنونی، مهاجمان می‌توانند داده‌هایی طراحی کنند که از شناسایی در فرآیندهای اعتبارسنجی دور بمانند و در عین حال به اهداف خود دست یابند. اگر تست شامل سناریوهایی که حملات مسمومیت هدفمند را پوشش می‌دهند نباشد، خطاهای واردشده ممکن است بدون شناسایی باقی بمانند.
- **محیط استقرار:**
  - اگر مهاجمان بتوانند بر داده‌هایی که در مرحله عملیاتی مدل پس از استقرار استفاده می‌شود تأثیر بگذارند، ممکن است رفتارهای مورد نظر را برای موارد خاص بدون تغییر مستقیم مدل ایجاد کنند.

۳-۸-۳ آسیب‌پذیری‌ها

حملات مسمومیت هدفمند از آسیب‌پذیری‌های خاص در خط لوله یادگیری ماشین بهره‌برداری می‌کنند تا رفتار مدل را به‌طور انتخابی دستکاری کنند. درک این آسیب‌پذیری‌ها برای توسعه دفاع‌های مؤثر بسیار مهم است. در زیر به آسیب‌پذیری‌های کلیدی که می‌توانند در حملات مسمومیت هدفمند مورد بهره‌برداری قرار گیرند، اشاره شده است:

- **کیفیت و یکپارچگی داده‌ها:**
  - عدم وجود فرآیندهای اعتبارسنجی دقیق می‌تواند اجازه دهد که داده‌های خراب یا دستکاری‌شده بدون شناسایی در آموزش مدل استفاده شوند و اتکا به داده‌های منابع تأییدنشده یا احتمالاً آسیب‌دیده، خطر مسمومیت داده‌ها را افزایش می‌دهد.
- **آسیب‌پذیری در فرآیند آموزشی:**
  - مدل‌هایی که به شدت به نقاط پرت و ناهنجاری‌ها در داده‌های آموزشی حساس هستند، بیشتر در معرض دستکاری هدفمند قرار دارند. الگوریتم‌هایی که به راحتی می‌توانند تحت تأثیر دستکاری داده‌های اقلیت یا نقاط خاص قرار گیرند، در معرض خطر هستند، زیرا مسمومیت هدفمند می‌تواند سوگیری‌های نظام‌مند ایجاد کند.
- **آسیب‌پذیری‌های معماری مدل:**
  - مدل‌های مستعد به بیش‌برازش می‌توانند توسط نمونه‌های هدفمندی که برای بهره‌برداری از این ضعف طراحی شده‌اند، فریب داده شوند و الگوهای نادرستی را یاد بگیرند که فقط به موارد هدفمند مربوط می‌شوند. عدم وجود روش‌های تنظیم‌گری مؤثر اجازه می‌دهد که داده‌های دستکاری‌شده تأثیر نامناسبی بر مدل داشته باشند.
- **مهندسی و انتخاب ویژگی:**
  - مهاجمان می‌توانند فرآیندهای استخراج یا انتخاب ویژگی را دستکاری کنند تا ویژگی‌های گمراه‌کننده‌ای را معرفی کنند که بر تفسیر مدل از ورودی‌های خاص تأثیر بگذارد. مدل‌هایی که به شدت به ویژگی‌های خاص وابسته هستند، ممکن است به دستکاری‌های هدفمند ویژگی‌ها آسیب‌پذیر باشند.
- **شکاف‌های آزمایش و اعتبارسنجی:**
  - پروتکل‌های آزمایش که شامل سناریوهای خصمانه یا موارد خاصی که احتمالاً هدف حملات قرار می‌گیرند، نیستند، مدل را آسیب‌پذیر می‌گذارند. عدم ارزیابی استحکام مدل در برابر داده‌های دستکاری‌شده می‌تواند اجازه دهد که داده‌های مسموم هدفمند اثرات مورد نظر

خود را بدون شناسایی اجرا کنند.

- **نقاط ضعف در خط لوله و فرآیندها:**
  - آسیب‌پذیری‌ها در پیش‌پردازش داده‌ها و تبدیل‌ها ممکن است برای معرفی سوگیری‌ها یا نادرستی‌های ظریف بدون شناسایی مورد بهره‌برداری قرار گیرند. بدون نظارت جامع بر خطوط لوله داده، دستکاری‌ها در جریان‌های ورودی داده ممکن است به موقع شناسایی نشوند.

۳-۸-۴ سناریو تهدید

در یک حمله مسمومیت هدفمند، یک مهاجم سعی دارد به‌طور ظریف داده‌های آموزشی را دستکاری کند تا مدل آموزش‌دیده برای ورودی‌های خاص اشتباه طبقه‌بندی کرده یا رفتار نامطلوبی از خود نشان دهد. این دستکاری بدون تأثیر قابل‌توجه بر دقت کلی مدل یا ایجاد سوءظن انجام می‌شود. این نوع حملات به‌ویژه در کاربردهای حساس مانند تشخیص تقلب، سامانه‌های امنیتی و وسایل نقلیه خودران خطرناک است؛ زیرا اشتباه طبقه‌بندی در موارد خاص می‌تواند عواقب جدی داشته باشد.

- **آماده‌سازی و شناسایی:**
  - درک مدل هدف: مهاجم مدل هدف، فرآیندهای آموزشی و منابع داده را تحلیل می‌کند تا آسیب‌پذیری‌های بالقوه و بردارهای حمله ارزشمند را شناسایی کند.
  - انتخاب ورودی‌های هدف: ورودی‌ها یا سناریوهای خاص بر اساس تأثیر یا ارزش بالای آن‌ها برای اهداف مهاجم انتخاب می‌شوند.
- **دستکاری داده:**
  - ایجاد نمونه‌های مسموم: مهاجم مجموعه کوچکی از نمونه‌های داده آموزشی را ایجاد یا اصلاح می‌کند. این نمونه‌ها به‌طور ظریف تغییر می‌کنند تا شامل برچسب‌های نادرست یا تغییرات ویژگی‌های سفارشی شوند که منجر به یادگیری الگوهای گمراه‌کننده توسط مدل می‌شود.
  - اطمینان از پنهان‌کاری: نمونه‌های طراحی‌شده به‌گونه‌ای ساخته می‌شوند که با داده‌های مشروع ترکیب شوند و خطر شناسایی در طول پیش‌پردازش داده و بازرسی‌های اولیه را کاهش دهند.
- **تزریق داده:**
  - مهاجم نمونه‌های مسموم را به مجموعه داده آموزشی تزریق می‌کند. این کار می‌تواند از طریق خطوط لوله داده آسیب‌دیده، آلودگی داده یا استفاده از دسترسی به مراحل جمع‌آوری

داده انجام شود.

- **آموزش مدل:**
  - مدل با استفاده از مجموعه داده کامل، شامل نمونه‌های مسموم، آموزش داده می‌شود. این نمونه‌ها به‌طور انتخابی فرآیند یادگیری را برای ورودی‌ها یا سناریوهای خاص منحرف می‌کنند.
- **پیاده‌سازی و فعال‌سازی:**
  - پس از پیاده‌سازی، مدل به‌طور طبیعی عمل می‌کند. با این حال، وقتی با ورودی‌های هدف‌گذاری‌شده مواجه می‌شود، رفتار نامطلوبی مانند اشتباه طبقه‌بندی یا تصمیم‌گیری نادرست را نشان می‌دهد.

۳-۸-۵ نشانه‌های احتمال نفوذ

حملات مسمومیت هدفمند می‌توانند مدل‌ها را وادار کنند تا پیش‌بینی‌ها یا طبقه‌بندی‌های نادرستی در مواجهه با ورودی‌های خاص انجام دهند.. در زیر به برخی از نشانه‌های احتمالی نفوذ رایج مرتبط با حملات مسمومیت هدفمند اشاره شده است:

- **داده‌های آموزشی غیرعادی:**
  - وجود نقاط پرت یا داده‌های غیرواقعی در مجموعه داده‌های آموزشی.
  - نقاط داده‌ای که به‌طور نظام‌مند پیش‌بینی‌های مدل را منحرف می‌کنند.
  - ناسازگاری‌های بین داده‌های آموزشی جدید و تاریخی.
- **رفتار غیرمنتظره مدل:**
  - کاهش ناگهانی دقت یا عملکرد مدل.
  - بیش‌برازش به موارد خاص در مجموعه داده.
  - اشتباه طبقه‌بندی ورودی‌های خاص، به‌ویژه آن‌هایی که هدف حمله هستند.
- **یکپارچگی منبع داده:**
  - عدم تأیید یا مشکوک بودن مبدا داده‌های آموزشی.
  - شواهدی از دستکاری در لاگ‌های داده یا تاریخچه نسخه‌ها.
  - عدم وجود فرآیندهای مناسب اعتبارسنجی و پاک‌سازی داده‌ها.
- **جابه‌جایی مدل:**
  - تغییر تدریجی در مرزهای تصمیم‌گیری مدل که با اهداف مسمومیت هم‌راستا است.
  - کاهش مداوم عملکرد مدل در طول زمان بدون تغییرات واضح در محیط.

- **رفتار آموزشی غیرعادی:**
  - زمان آموزش طولانی‌تر بدون توضیح منطقی.
  - انحراف در معیارهای عملکرد آموزشی و اعتبارسنجی.
- **شاخص‌های خارجی:**
  - گزارش‌هایی از حملات مشابه در صنعت یا بخش مربوطه.
  - دشمنان شناخته‌شده که سازمان را با تهدیدات پیشرفته و مداوم هدف قرار می‌دهند.
- **ناهنجاری‌های دسترسی:**
  - الگوهای دسترسی غیرمجاز یا غیرعادی به محیط آموزشی.
  - تغییرات در مجوزهای دسترسی برای مجموعه داده‌ها یا مدل‌های کلیدی.
- **حلقه‌های بازخورد:**
  - ورودی‌های تکراری با بازخورد منفی از کاربران نهایی یا سامانه‌ها.
  - شکایات مکرر درباره خروجی‌های خاص مدل که با اهداف مسمومیت هم‌راستا است.

۳-۸-۶ تأثیرات

حملات مسمومیت هدفمند بر روی مدل‌های یادگیری ماشین می‌توانند تأثیرات متنوعی داشته باشند، که بستگی به اهداف مهاجم و کاربرد مدل تحت تأثیر دارد. در زیر برخی از تأثیرات بالقوه این حملات آورده شده است:

- **کاهش عملکرد مدل:**
  - مستقیم‌ترین تأثیر، کاهش دقت مدل، به‌ویژه برای کلاس‌ها یا ورودی‌های هدفمند خاص است. مدل ممکن است به‌طور مداوم ورودی‌های خاصی را اشتباه طبقه‌بندی کند که منجر به خروجی‌های نادرست می‌شود.
- **نقض‌های امنیتی:**
  - اگر مدلی برای مقاصد امنیتی، مانند تشخیص نفوذ، استفاده شود، وضعیت تخریب‌شده آن می‌تواند منجر به نقض‌های امنیتی غیرقابل شناسایی شود. مهاجمان ممکن است از اشتباهات طبقه‌بندی برای دستیابی غیرمجاز به سامانه‌ها یا داده‌های حساس سوءاستفاده کنند.
- **زیان مالی:**
  - تصمیم‌گیری نادرست می‌تواند منجر به زیان‌های مالی مستقیم، مانند پیش‌بینی‌های مالی

نادرست یا معاملات تقلبی شود. عملیات تجاری که به خروجی‌های دقیق مدل وابسته هستند، ممکن است ناکارآمد یا ناکام شوند که منجر به عواقب مالی غیرمستقیم خواهد شد.

- **آسیب به اعتماد و شهرت:**
  - کاربران و ذینفعان ممکن است اعتماد خود را به اعتبار و یکپارچگی سامانه از دست بدهند. مشکلات طولانی‌مدت یا اطلاع عمومی از یک سامانه تخریب‌شده می‌تواند به شهرت سازمان آسیب برساند.
- **مسائل قانونی و انطباق:**
  - نهادهای نظارتی ممکن است جریمه‌هایی را در صورت شناسایی ناکارآمدی در شیوه‌های مدیریت و حفاظت از داده‌ها اعمال کنند و یا کاربران تحت تأثیر تصمیمات تخریب‌شده ممکن است به دنبال جبران قانونی باشند.
- **اختلال در عملیات:**
  - تنظیمات برای کاهش اثرات حمله، مانند بازآموزی مدل‌ها، ممکن است باعث اختلال در عملیات عادی شود. ممکن است منابع و زمان اضافی برای شناسایی، تجزیه و تحلیل و کاهش اثرات داده‌های مسموم مورد نیاز باشد.
- **از دست دادن مزیت راهبردی:**
  - رقبا ممکن است در صورتی که تصمیمات راهبردی یک کسب‌وکار، که توسط بینش‌های یادگیری ماشین تقویت شده است، تحت تأثیر قرار گیرد، مزیتی کسب کنند. علاوه بر این ترس از حملات بیشتر ممکن است منجر به کاهش نوآوری، به‌ویژه در پذیرش مدل‌های پیشرفته یادگیری ماشین شود.
- **تخریب تجربه کاربری:**
  - کاربران نهایی ممکن است با کیفیت خدمات کمتری مواجه شوند، به‌ویژه اگر مدل در ارائه خدمات شخصی‌سازی‌شده حیاتی باشد.
- **انتشار خروجی‌های نادرست:**
  - اگر خروجی‌های مدل در سامانه‌ها یا خطوط لوله بزرگ‌تر ادغام شوند، پیش‌بینی‌های نادرست می‌توانند گسترش یابند و مشکلات سامانهیک وسیع‌تری ایجاد کنند.

۳-۸-۷ راه‌های مقابله

کاهش حملات مسمومیت هدفمند در یادگیری ماشین شامل ترکیبی از اقدامات پیشگیرانه در طول

جمع‌آوری داده، آموزش مدل و پیاده‌سازی است. در زیر به برخی از راهبردها برای کاهش این حملات اشاره شده است:

- **اعتبارسنجی داده قوی:**
  - پیاده‌سازی پروتکل‌های دقیق اعتبارسنجی و پاک‌سازی داده تا نقاط داده ناهنجار قبل از استفاده در آموزش شناسایی و حذف شوند. و استفاده از تکنیک‌های تشخیص ناهنجاری آماری برای شناسایی نقاط داده‌ای که به‌طور قابل توجهی از توزیع‌های مورد انتظار منحرف شده‌اند.
- **ردیابی منشأ و یکپارچگی داده:**
  - ردیابی منشأ داده برای حصول اطمینان از اعتبار و تأیید منبع و تاریخچه نقاط داده و استفاده از تکنیک‌های رمزنگاری برای تأمین یکپارچگی داده‌ها در طول جمع‌آوری، انتقال و ذخیره‌سازی.
- **افزونگی و اعتبارسنجی متقابل:**
  - استفاده از اعتبارسنجی متقابل و روش‌های مجموعه‌ای برای اطمینان از استحکام مدل با آموزش بر روی چندین زیرمجموعه داده و مقایسه نتایج برای یکپارچگی. علاوه بر این آموزش مدل‌ها بر روی داده‌های چندین منبع تا ریسک تأثیر یک مجموعه داده مخرب بر مدل کلی کاهش یابد.
- **آموزش خصمانه:**
  - به‌کارگیری تکنیک‌های آموزش خصمانه برای مقاوم‌سازی مدل‌ها در برابر داده‌های آلوده بالقوه در طول مرحله آموزش و کمک به آن‌ها تا یاد بگیرند چگونه به این دستکاری‌ها مقاومت کنند.
- **یادگیری تدریجی و آنلاین:**
  - پیاده‌سازی رویکردهای یادگیری تدریجی که مدل‌ها را به‌طور مکرر و در مراحل کوچکتر به‌روزرسانی کنند، که امکان نظارت و کاهش هرگونه تغییر ناگهانی در رفتار مدل را فراهم می‌کند.
- **بررسی و آزمایش منظم مدل:**
  - به‌طور منظم مدل‌ها برای سوگیری‌ها یا رفتارهای غیرمنتظره بررسی و تست شوند و از مجموعه داده‌های تستی که برای نشان دادن اثرات مسمومیت هدفمند طراحی شده‌اند، استفاده شود. هم‌چنین پیاده‌سازی چرخه‌های آموزشی و ارزیابی دوره‌ای تا آسیب‌پذیری‌های

بالقوه ناشی از داده‌های آلوده شناسایی و اصلاح شوند نیز می‌تواند کمک‌کننده باشد.

- **تنوع منابع داده‌ای:**
  - منابع داده باید متنوع شوند تا وابستگی به هر منبع داده واحدی که ممکن است هدف حمله قرار گیرد، کاهش یابد. علاوه بر این از تکنیک‌های تقویت داده‌ها برای بهبود مصنوعی مجموعه‌های داده استفاده و تنوع و استحکام را به مواد آموزشی اضافه شوند.
- **تشخیص ناهنجاری و نظارت:**
  - پیاده‌سازی سامانه‌های تشخیص ناهنجاری در زمان واقعی را برای نظارت بر خروجی‌های مدل پیاده‌سازی شده و برای هر الگوی غیرمعمول یا رفتاری که ممکن است نشان‌دهنده داده‌های آموزشی دستکاری‌شده باشد، هشدار داده شود. هم‌چنین می‌توان آستانه‌هایی برای انحرافات قابل قبول در عملکرد مدل تعیین گرد و در صورت تجاوز از آن‌ها، تحقیقات بیشتری آغاز شود.
- **کنترل‌های دسترسی و امنیت:**
  - کنترل‌های دسترسی سختگیرانه و تدابیر احراز هویت برای محافظت از ورودی‌های داده و پارامترهای مدل در برابر تغییرات غیرمجاز پیاده‌سازی شود. علاوه بر این گزارش‌های دسترسی به داده‌ها برای فعالیت‌های مشکوک که ممکن است نشان‌دهنده تلاشی برای معرفی داده‌های آلوده باشد، نظارت شود.
- **احتیاط در همکاری و جمع‌سپاری:**
  - هنگام استفاده از داده‌های جمع‌سپاری یا ورودی‌های پلتفرم‌های همکاری، احتیاط شود، زیرا این‌ها معمولاً راه‌هایی رایج برای مسمومیت هدفمند هستند.

۳-۹ مسمومیت با درب پشتی

حملات مسمومیت با درب پشتی[1] نوعی حمله خصمانه به مدل‌های یادگیری ماشین هستند که در آن یک مهاجم به‌طور ظریف داده‌های آموزشی را دستکاری می‌کند تا یک "درب پشتی" را در مدل مستقر کند. این درب پشتی یک الگوی ورودی خاص یا محرکی است که هنگامی که در داده‌های ورودی وجود دارد، باعث می‌شود مدل به روشی که توسط مهاجم انتخاب شده است عمل کند و اغلب پیش‌بینی‌های نادرست یا از پیش تعیین‌شده‌ای را تولید کند.

۳-۹-۱ ویژگی‌های کلیدی

حملات مسمومیت با درب پشتی ویژگی‌های کلیدی متعددی دارند که آن‌ها را از سایر انواع حملات

Targeted Poisoning 1. Backdoor Poisoning

خصمانه متمایز می‌کند. این ویژگی‌ها شامل موارد زیر است:

- **فعال‌سازی مبتنی بر محرک:**
  - درب پشتی در مدل، با یک الگوی محرک خاص و از پیش تعیین‌شده که مهاجم در داده‌ها در مرحله آموزش جایگذاری کرده، فعال می‌شود. این محرک می‌تواند یک الگوی بصری، مقدار داده خاص یا دنباله‌ای باشد که به نوع داده‌ای که مدل پردازش می‌کند بستگی دارد.
- **پنهان بودن و ظرافت:**
  - تغییرات در داده‌های آموزشی یا مدل حداقلی است و به‌گونه‌ای راهبردی طراحی شده تا از تشخیص جلوگیری کند. مدل دارای درب پشتی در شرایط عادی مورد انتظار عمل می‌کند و فقط زمانی که محرک وجود دارد، رفتار نادرستی از خود نشان می‌دهد.
- **عملکرد طبیعی بر روی داده‌های تمیز:**
  - مدل به عملکرد صحیح خود ادامه می‌دهد و به‌خوبی بر روی ورودی‌های بی‌خطر که محرک را شامل نمی‌شوند، عمل می‌کند و دقت و رفتار مورد انتظار خود را در اکثر شرایط حفظ می‌کند.
- **نادرستی انتخابی:**
  - تنها ورودی‌هایی که شامل محرک هستند، منجر به نادرستی یا خروجی مورد نظر مهاجم می‌شوند. این انتخاب‌پذیری شناسایی حمله را از طریق روش‌های معمول آزمایش دشوار می‌کند.
- **سهولت درج:**
  - مهاجم معمولاً فقط نیاز دارد که بخش کوچکی از مجموعه داده را تغییر دهد یا چند نمونه مخرب با محرک را معرفی کند، که اجرای حمله را نسبتاً آسان می‌سازد، به‌ویژه اگر آن‌ها کنترل یا تأثیر جزئی بر داده‌های آموزشی داشته باشند.
- **پایداری:**
  - پس از ایجاد، درب پشتی تمایل به ماندگاری در طول به‌روزرسانی‌های مدل یا تنظیمات دقیق دارد، مگر اینکه تدابیر خاصی برای شناسایی و حذف آن اتخاذ شود، که این موضوع آن را به یک تهدید پایدار تبدیل می‌کند.
- **قابلیت کاربرد وسیع:**
  - حملات درب پشتی می‌توانند در طیف وسیعی از مدل‌های یادگیری ماشین و حوزه‌ها، از جمله بینایی کامپیوتری، پردازش زبان طبیعی و حتی یادگیری تقویتی اجرا شوند.

۳-۹-۲ دارایی‌های هدف

در زمینه حملات مسمومیت با درب پشتی، دارایی‌های هدف به عناصری در یک خط لوله یادگیری ماشین اشاره دارد که می‌توانند توسط مهاجم مورد تعرض یا دستکاری قرار گیرند. این دارایی‌ها شامل موارد زیر است:

- **داده‌های آموزشی:**
  - مهاجمین به دنبال مسموم کردن مجموعه داده‌های آموزشی با وارد کردن ورودی‌های مخرب هستند که شامل مژحرک درب پشتی می‌باشند. به خطر انداختن یکپارچگی داده‌ها به مهاجمین این امکان را می‌دهد که محرک‌ها را بدون شناسایی جایگذاری کنند. مهاجمین ممکن است به منابع یا کانال‌هایی که از طریق آن‌ها داده‌های آموزشی به‌دست می‌آید، هدف قرار دهند، به‌ویژه اگر این منابع به‌صورت عمومی قابل دسترسی یا به‌خوبی مدیریت نشده باشند.
- **مدل یادگیری ماشین:**
  - با جایگذاری کردن یک درب پشتی، مهاجمین توانایی مدل را برای انجام پیش‌بینی‌های قابل اعتماد تحت شرایط خاص به خطر می‌اندازند، معمولاً در ارتباط با وجود محرک. در طول به‌روزرسانی‌ها یا دوباره‌آموزی، مهاجمین ممکن است این مرحله را هدف قرار دهند تا درب‌های پشتی را جایگذاری یا حفظ کنند، به‌ویژه اگر این فرآیندها بدون پروتکل‌های تأیید قوی انجام شوند.
- **فرآیند استنتاج:**
  - هدف اصلی یک درب پشتی تأثیرگذاری بر پیش‌بینی‌های مدل تحت شرایط خاص است در حالی که عملکرد معمولی در سایر شرایط حفظ می‌شود. درب‌های پشتی در مرحله استنتاج با ایجاد پیش‌بینی‌های نادرست زمانی که ورودی شامل محرک است، از این مرحله سوءاستفاده می‌کنند و ممکن است بر تصمیم‌گیری سامانه تأثیر بگذارند.
- **خطوط لوله داده:**
  - دستکاری مراحل آماده‌سازی داده برای مدل می‌تواند به درج داده‌های محرک به شیوه‌هایی که به‌راحتی شناسایی نمی‌شوند، کمک کند. از آنجایی که داده‌های برچسب‌گذاری شده برای یادگیری نظارت شده حیاتی هستند، خطاها در برچسب‌گذاری می‌توانند توسط مهاجمین مورد سوءاستفاده قرار گیرند تا فرآیند جایگذاری کردن درب پشتی را تسهیل کنند.
- **محیط استقرار:**
  - مهاجمین از اعتماد به خروجی‌های یادگیری ماشین سوءاستفاده می‌کنند و باعث می‌شوند

سامانه به‌طور غیرمنتظره‌ای عمل کند که ممکن است به نقض‌های امنیتی یا اختلالات عملکردی منجر شود، به‌ویژه در کاربردهای حیاتی. در زمینه‌هایی که قابلیت اطمینان حیاتی است (مانند خدمات مالی و بهداشت و درمان)، یک مدل دارای درب پشتی می‌تواند به اعتبار کاربر و سامانه آسیب بزند وقتی که باعث تصمیم‌گیری‌های نادرست می‌شود.

۳-۹-۳ آسیب‌پذیری‌ها

حملات مسمومیت با درب پشتی از آسیب‌پذیری‌های خاصی در خط لوله یادگیری ماشین سوءاستفاده می‌کنند تا درب‌های پشتی را وارد و فعال کنند. درک این آسیب‌پذیری‌ها برای توسعه سازوکارهای دفاعی مؤثر حیاتی است. در ادامه، برخی از کلیدی‌ترین آسیب‌پذیری‌هایی که می‌توانند در حملات مسمومیت با درب پشتی هدف قرار گیرند، آورده شده است:

- **یکپارچگی و امنیت داده:**
  - اگر داده‌های آموزشی از منابع غیرقابل اعتماد یا ناامن جمع‌آوری شوند، ممکن است برای یک مهاجم آسان باشد که داده‌های مسموم را وارد کند. عدم وجود مراحل پیش‌پردازش قوی می‌تواند به نمونه‌های فعال‌شده توسط درب پشتی اجازه دهد به‌راحتی در مجموعه داده‌های آموزشی ادغام شوند.
- **عدم اعتبارسنجی داده:**
  - عدم اجرای اقدامات اعتبارسنجی و شناسایی ناهنجاری دقیق برای داده‌های آموزشی اجازه می‌دهد تا ورودی‌های مخرب بدون شناسایی وارد شوند. جمع‌آوری داده‌های بیشتر بدون اطمینان از کیفیت آن‌ها می‌تواند منجر به آسیب‌پذیری‌هایی شود که داده‌های مسموم در آموزش استفاده می‌شوند.
- **فرآیند آموزش مدل:**
  - بسیاری از فرآیندهای آموزشی شامل بررسی‌هایی برای مقاومت در برابر دستکاری‌های خصمانه نیستند و این موضوع آن‌ها را در برابر درج درب پشتی آسیب‌پذیر می‌کند. مدل‌هایی که بر روی مجموعه‌های داده بسیار بزرگ آموزش می‌بینند ممکن است نتوانند چند نمونه مسموم را به دلیل فیلتر کردن بر اساس اهمیت آماری شناسایی کنند.
- **معماری مدل:**
  - مدل‌هایی که تمایل به سازگاری زیاد با داده‌های آموزشی دارند ممکن است الگوها یا محرک‌های نادری را که توسط مهاجمان معرفی شده‌اند، قوی‌تر شناسایی و واکنش نشان دهند. مدل‌های بسیار پیچیده ممکن است به‌طور ناخواسته همبستگی‌های کاذب مانند

مواردی که توسط الگوهای درب پشتی ارائه می‌شوند را یاد بگیرند.

- **زیرساخت یادگیری ماشین:**
  - عدم وجود نظارت و بازرسی در خط لوله یادگیری ماشین از جمع‌آوری داده تا استقرار می‌تواند از شناسایی زودهنگام فعالیت‌های درب پشتی جلوگیری کند. هم‌چنین عدم پیگیری و مقایسه نظام‌مند نسخه‌های مختلف مدل می‌تواند منجر به آسیب‌پذیری‌های غیرقابل شناسایی شود که از مراحل قبلی منتقل شده‌اند.
- **شیوه‌های استقرار:**
  - رویه‌های آزمایشی ناکافی ممکن است موارد حاشیه‌ای یا سناریوهای خاصی که در آن‌ها درب پشتی به احتمال زیاد فعال می‌شود را پوشش ندهد. استفاده از مدل‌های پیش‌آموزش‌دیده یا داده‌های شخص ثالث بدون اعتبارسنجی دقیق، سامانه را در برابر درب پشتی‌های به ارث‌رسیده آسیب‌پذیر می‌سازد.
- **آگاهی محدود امنیتی:**
  - تیم‌ها ممکن است از مهارت‌ها یا آگاهی‌های لازم برای شناسایی و رسیدگی به راهبردهای بالقوه درب پشتی برخوردار نباشند. هم‌چنین برخی سازمان‌ها ممکن است سطح تهدید ناشی از حملات درب پشتی را به‌طور کافی ارزیابی نکنند و این موضوع منجر به تدابیر امنیتی ناکافی شود.

۴-۹-۳ سناریو تهدید

در یک حمله مسمومیت با درب پشتی، دشمن به دنبال جایگذاری کردن یک آسیب‌پذیری پنهان (درب پشتی) در یک مدل یادگیری ماشین است. این موضوع به مدل اجازه می‌دهد که تحت کنترل یا سوءاستفاده قرار گیرد زمانی که ورودی‌های خاصی، که به آن‌ها محرک گفته می‌شود، ارائه می‌شود و در مواقع دیگر به‌طور عادی عمل کند. این سناریو به‌خصوص در کاربردهایی با الزامات امنیتی یا ایمنی بالا، مانند وسایل نقلیه خودران، سامانه‌های شناسایی چهره و نظارت بر زیرساخت‌های حیاتی، تهدیدآمیز است، جایی که نادرستی‌ها می‌تواند به عواقب شدید منجر شود.

- **آماده‌سازی و دسترسی:**
  - دست‌یابی: مهاجم به خط لوله آموزشی دسترسی پیدا می‌کند. این ممکن است از طریق دسترسی مستقیم به داده‌ها یا تأثیرات غیرمستقیم، مانند مشارکت در داده‌های یک مجموعه داده عمومی، اتفاق بیفتد.
  - طراحی محرک: مهاجم یک الگوی محرک طراحی می‌کند که درب پشتی را فعال کند. این

محرک باید متمایز اما به‌قدری ظریف باشد که در عملیات عادی و تست‌ها شناسایی نشود.

- **مسمومیت داده:**
  - وارد کردن نمونه‌های مسموم: مهاجم تعداد کمی از نمونه‌های آموزشی را وارد می‌کند که شامل الگوی محرک هستند و این‌ها را با برچسب نادرست مورد نظر مرتبط می‌کند. این کار به‌گونه‌ای انجام می‌شود که تأثیر قابل توجهی بر عملکرد مدل در ورودی‌های بدون محرک نداشته باشد.
- **آموزش مدل:**
  - آموزش با داده‌های مسموم: فرآیند آموزشی کامل ادامه می‌یابد و هم داده‌های تمیز و هم داده‌های مسموم را در بر می‌گیرد. مدل رفتار صحیح را برای بیشتر داده‌ها یاد می‌گیرد و به‌طور ناخواسته شرایط فعال‌سازی درب پشتی را برای ورودی‌های حاوی محرک یاد می‌گیرد.
- **تأیید و استقرار:**
  - آزمایش و اعتبارسنجی: رویه‌های آزمایشی استاندارد معمولاً ورودی‌های محرک درب پشتی را شامل نمی‌شوند و اجازه می‌دهند مدل مسموم بدون شناسایی در آزمون‌ها موفق عمل کند.
  - استقرار: مدل به محیط تولید منتقل می‌شود و در آن‌جا به‌طور عادی عمل می‌کند تا زمانی که محرک به آن معرفی شود.
- **سوءاستفاده:**
  - فعال‌سازی: مهاجم ورودی‌هایی حاوی محرک را به مدل معرفی می‌کند که باعث می‌شود مدل به‌طور غیرمنتظره‌ای عمل کند. نتیجه می‌تواند از نادرستی‌های ساده تا عواقب سامانهی متفاوت باشد بسته به کاربرد مدل.

۳-۹-۵ نشانه‌های احتمال نفوذ

نشانه‌های احتمال نفوذ برای حملات مسموم‌سازی درب پشتی به نشانه‌ها یا شواهدی اشاره دارند که نشان می‌دهند یک مدل تحت چنین حمله‌ای قرار گرفته است. شناسایی این شاخص‌ها می‌تواند چالش‌برانگیز باشد به دلیل ماهیت پنهانی حملات درب پشتی، اما چندین نشانه‌های احتمال نفوذ بالقوه می‌تواند در شناسایی مدل‌های آسیب‌دیده کمک کند:

- **رفتار غیرعادی مدل:**
  - مدل به طور مداوم ورودی‌های خاصی را به اشتباه طبقه‌بندی یا پیش‌بینی‌های نادرستی انجام می‌دهد. هم‌چنین مدل ممکن است تصمیمات یا خروجی‌هایی تولید کند که با منطق کسب‌وکار یا انتظارات هم‌راستا نیست، به ویژه در سناریوهایی که در طول اعتبارسنجی مشاهده نشده‌اند.

- **نقص‌های عملکرد مدل:**
  - آزمایش‌ها نشان می‌دهد که فقط در حضور ورودی‌های خاص، عملکرد به طور قابل توجهی کاهش می‌یابد که می‌تواند به رفتارهای مبتنی بر محرک اشاره کند. تفاوت میان نتایج اعتبارسنجی و عملکرد در سناریوهای واقعی می‌تواند نشان‌دهنده دستکاری باشد، به ویژه اگر چنین اختلافاتی تحت شرایط خاصی رخ دهد.
- **ناهنجاری‌ها در داده‌های آموزشی:**
  - وجود الگوهای غیرمعمول یا غیرمنتظره در مجموعه داده که می‌تواند به محرک‌های بالقوه‌ای برای درب پشتی مربوط باشد. علاوه بر این نشانه‌هایی که نشان می‌دهد داده‌ها دستکاری شده‌اند، از جمله برچسب‌های غیرمنتظره، داده‌های تکراری با تغییرات، یا داده‌های به‌دست‌آمده از منابع ناشناخته.
- **گزارش‌های امنیتی و الگوهای دسترسی:**
  - شناسایی دسترسی یا تغییرات غیرمجاز در داده‌های آموزشی یا دارایی‌های مدل می‌تواند نشانه‌ای از تلاش‌های تزریق باشد. علاوه بر این گزارش‌هایی که نشان‌دهنده تغییرات در خط لوله پردازش داده بدون توجیه شناخته‌شده هستند، می‌تواند به تلاش‌های مسموم‌سازی داده اشاره کند.
- **آزمون‌ها و ممیزی‌های مدل:**
  - آزمون‌های خاصی می‌توانند طراحی شوند تا برای شناسایی فعال‌سازی‌های درب پشتی با معرفی الگوهای ورودی متنوع در طول آزمون‌ها بررسی شوند. سامانه‌های نظارتی ممکن است ناهنجاری‌هایی را در نحوه پردازش مدل نسبت به ورودی‌های خاص شناسایی کنند، از جمله تأخیرها یا خطاهایی که با داده‌های دیگر دیده نمی‌شود.
- **بازخورد:**
  - بازخورد کاربران که نشان‌دهنده پیش‌بینی‌های عجیب یا نادرست در زمینه‌های خاص است می‌تواند به فعال شدن محرک‌های درب پشتی اشاره کند. نتایج عملیاتی غیرمنتظره و وجود ناهنجاری در نتایج، به ویژه در سامانه‌هایی که در آن‌ها مدل‌های یادگیری ماشین تصمیم‌گیری می‌کنند، می‌تواند نشانه‌ای از مدل آسیب‌دیده باشد.

۳-۹-۶ تأثیرات

حملات مسمومیت درب پشتی می‌توانند تأثیرات قابل توجه و بلندمدتی بر سامانه‌های یادگیری ماشین و محیط‌هایی که در آن‌ها پیاده‌سازی می‌شوند، داشته باشند. در ادامه برخی از تأثیرات کلیدی

مرتبط با این حملات آورده شده است:

- **نقض‌های امنیتی:**
  - درب‌های پشتی می‌توانند به افراد غیرمجاز اجازه دهند تا با بهره‌برداری از شرایط محرک، سامانه را دستکاری کنند و احتمالاً به افشای داده‌های حساس یا دسترسی غیرمجاز به ویژگی‌های محدود منجر شوند. مهاجمان می‌توانند از مدل برای انجام اقداماتی که با اهداف آن‌ها هم‌راستا است، سوءاستفاده کنند که ممکن است فعالیت‌های مخرب مختلفی مانند کلاهبرداری یا خرابکاری را تسهیل کند.
- **اختلال در عملیات:**
  - فعال شدن یک درب پشتی می‌تواند به خروجی‌های غیرقابل پیش‌بینی یا نادرست منجر شود که به شدت بر قابلیت اعتماد و اعتبار سامانه تأثیر می‌گذارد، به ویژه در برنامه‌های حیاتی مانند وسایل نقلیه خودران یا تشخیص پزشکی. شناسایی و اصلاح وجود یک درب پشتی ممکن است نیاز به توقف قابل توجه و تخصیص منابع برای تحلیل داشته باشد که به اختلال عملیاتی و هزینه‌های بازیابی قابل توجهی منجر می‌شود.
- **آسیب به شهرت و اعتماد:**
  - ذینفعان ممکن است اعتماد خود را به سازمان یا تأمین‌کننده سامانه از دست بدهند اگر یکپارچگی سامانه‌های یادگیری ماشین آن‌ها به خطر بیفتد، که بر روابط مشتری و اعتبار بازار تأثیر می‌گذارد. نگرانی‌های مداوم امنیتی می‌تواند به شهرت سازمان‌ها آسیب بزند، به ویژه در بخش‌هایی که اعتماد و قابلیت اعتماد اهمیت بالایی دارد (مانند مالی و بهداشت).
- **عواقب مالی و اقتصادی:**
  - سوءاستفاده از سامانه‌ها به دلیل یک درب پشتی می‌تواند به ضررهای مالی مستقیم منجر شود اگر حمله منجر به کلاهبرداری، سرقت یا کاهش کارایی عملیاتی شود. سازمان‌ها ممکن است به دلیل نیاز به تدابیر امنیتی بهبود یافته، تحلیل‌های جنایی و سرمایه‌گذاری در زیرساخت‌های یادگیری ماشین قوی‌تر، هزینه‌های بیشتری متحمل شوند.
- **پیامدهای قانونی و نظارتی:**
  - حملات درب پشتی می‌تواند به نقض‌های مقررات حفاظت از داده‌ها و الزامات انطباق منجر شود که ممکن است به جریمه‌ها و مجازات‌های قانونی بینجامد. سازمان‌ها ممکن است با دعاوی قانونی از سوی طرف‌های آسیب‌دیده مواجه شوند اگر ثابت شود که آن‌ها در پیشگیری یا پاسخ مناسب به یک حمله ناکام بوده‌اند.

- **تأثیرات بر کاربران:**
  - کاربران یا مشتریانی که به سامانه یادگیری ماشین برای تصمیم‌گیری تکیه می‌کنند ممکن است اطلاعات گمراه‌کننده یا نادرستی دریافت کنند که می‌تواند به نتایج مضر منجر شود. فعال شدن یک درب پشتی ممکن است حریم خصوصی کاربران را به خطر بیندازد و منجر به دسترسی یا افشای غیرمجاز داده‌ها شود.
- **تخریب بلندمدت یکپارچگی سامانه:**
  - با گذشت زمان، فعال شدن یک درب پشتی ممکن است به خطاهای تجمعی منجر شود یا عملکرد کلی مدل را کاهش دهد که نیاز به آموزش دوباره یا تعویض مدل‌های آسیب‌دیده را ایجاب می‌کند.

۳-۹-۷ راه‌های مقابله

کاهش حملات مسمومیت با درب پشتی شامل ترکیبی از راهبردها است که به شناسایی و پیشگیری از دستکاری غیرمجاز داده‌ها، اطمینان از آموزش قوی مدل و حفظ یکپارچگی کلی سامانه می‌پردازد. در ادامه برخی از اقدامات کلیدی کاهش خطر آمده است:

- **مدیریت و اعتبارسنجی داده‌ها:**
  - اطمینان حاصل شود که داده‌های آموزشی از منابع معتبر و قابل اعتماد تأمین می‌شود. همیشه باید به داده‌های ناشناخته یا غیرقابل اعتماد مشکوک بود. باید راهبردهایی برای پاک‌سازی و نرمال‌سازی داده‌ها قبل از استفاده پیاده‌سازی شود و داده‌های پرت یا عناصر مشکوک که ممکن است نشانه‌ای از محرک‌های تزریق‌شده باشد، حذف شوند.
- **تکنیک‌های آموزشی قوی:**
  - نمونه‌های خصمانه و محرک‌های درب پشتی در فرآیند آموزش گنجانده شود تا توانایی مدل برای شناسایی و خنثی‌سازی تهدیدات بالقوه افزایش یابد. از تکنیک‌های حریم خصوصی تفاضلی برای محدود کردن تأثیر نمونه‌های آموزشی فردی استفاده شده و خطر تغییر قابل توجه رفتار مدل به وسیله الگوهای درب پشتی کاهش داده شود.
- **ارزیابی دقیق مدل:**
  - سامانه‌های شناسایی ناهنجاری مستقر شود که بتوانند الگوهای پیش‌بینی غیرعادی را هنگام مواجهه با ورودی‌های غیرمنتظره شناسایی کنند. مجموعه‌های آزمایشی توسعه داده شود که شامل دامنه‌ای از سناریوهای خصمانه و مرتبط با درب پشتی باشد تا استحکام مدل در برابر دستکاری‌های هدفمند ارزیابی شود.

- **نظارت و ممیزی مدل:**
  - نظارت لحظه‌ای بر خروجی‌های مدل پیاده‌سازی شده تا ناهنجاری‌ها در پیش‌بینی‌ها، به ویژه آن‌هایی که با محرک‌های بالقوه هم‌راستا هستند، شناسایی شود. ممیزی‌های منظم از عملکرد و رفتار مدل انجام شود و برای اختلافاتی که می‌تواند نشانه‌ای از فعال شدن درب پشتی باشد، بررسی‌های لازم صورت گیرد.
- **کنترل دسترسی و امنیت:**
  - دسترسی به داده‌های آموزشی و مدل‌ها تنها به افراد مجاز محدود کنید و از روش‌های احراز هویت قوی و کنترل‌های دسترسی استفاده کنید. گزارش‌های دقیقی از دسترسی و تغییرات داده‌ها نگه‌داری شده تا قابلیت ردیابی را فراهم کرده و نقاط احتمالی نفوذ شناسایی شود.
- **تنوع در معماری مدل:**
  - از روش‌های مجموعه‌ای در مدل‌ها استفاده شده تا تأثیر هر مدل آسیب‌دیده را کاهش داده و موفقیت حمله درب پشتی را در تمام مدل‌ها دشوارتر شود. از پیچیدگی بیش از حد مدل اجتناب شده که ممکن است به طور ناخواسته همبستگی‌های مضر، مانند محرک‌های درب پشتی را جذب کند.
- **به‌روزرسانی‌های منظم و مدیریت وصله[1]:**
  - مدل‌ها به طور منظم با داده‌های به‌روز و معتبر آموزش مجدد شود تا خطر تهدیدات مداوم کاهش یابد. باید با آخرین به‌روزرسانی‌ها در چارچوب‌های یادگیری ماشین و کتابخانه‌ها که ممکن است شامل وصله‌های امنیتی برای آسیب‌پذیری‌های شناخته‌شده باشد، به‌روز ماند.
- **تعامل با جامعه و اطلاعات تهدید:**
  - با جامعه امنیت یادگیری ماشین همکاری شود تا اطلاعات مربوط به تهدیدات نوظهور و تکنیک‌های کاهش خطر به اشتراک گذاشته شود. باید از طریق تعامل فعال با گروه‌های تحقیقاتی و صنعتی، از به‌روزرسانی‌ها و تکنیک‌های جدید در حملات خصمانه و دفاع‌ها مطلع شد.

۳-۱۰ مسمومیت در دسترس‌پذیری

حملات مسمومیت دسترس‌پذیری[2] در زمینه‌های یادگیری ماشین به اقدام‌های خصمانه‌ای اشاره دارد که هدف آن‌ها کاهش دسترسی و عملکرد یک مدل یا سامانه یادگیری ماشین است. هدف این

1. Patch
2. Availability Poisoning

حملات این است که مدل را غیرقابل اعتماد یا غیرقابل استفاده کنند و بدین ترتیب دسترسی آن را برای برنامه‌ها یا خدمات مورد نظر تحت تأثیر قرار دهند.

۳-۱۰-۱ ویژگی‌های کلیدی

حملات مسمومیت دسترس‌پذیری دارای چندین ویژگی کلیدی هستند که روش‌هایی را که این حملات می‌توانند عملکرد و دسترسی سامانه را کاهش دهند، برجسته می‌سازند:

- **کاهش عملکرد:**
  - نتیجه اصلی این حملات، کاهش قابل توجه عملکرد مدل یادگیری ماشین است که منجر به افزایش نرخ خطا و پیش‌بینی‌های غیرقابل اعتماد در دامنه وسیعی از ورودی‌ها می‌شود.
- **تأثیر گسترده:**
  - بر خلاف حملات هدفمند، مسمومیت دسترس‌پذیری به خروجی‌های مدل به طور گسترده تأثیر می‌گذارد و توانایی کلی آن برای عملکرد صحیح را مختل می‌کند و نه فقط در شرایط محرک خاص.
- **دستکاری داده:**
  - مهاجمان داده‌های خراب، دارای نویز یا تغییر یافته خصمانه را به مجموعه داده‌های آموزشی وارد می‌کنند. این داده‌ها ممکن است شامل برچسب‌های نادرست، ویژگی‌های غیرمربوط یا ناهنجاری‌هایی باشد که فرآیند یادگیری را گیج می‌کنند.
- **پنهان‌سازی و اجتناب:**
  - داده‌های مخرب غالباً طوری طراحی می‌شوند که با داده‌های مشروع ترکیب شوند، که شناسایی آن‌ها بدون بررسی دقیق دشوار است. این باعث می‌شود داده‌های مسموم در بررسی‌های استاندارد اعتبارسنجی پنهان بمانند.
- **هدف غیرقابل استفاده کردن:**
  - هدف نهایی این است که مدل یادگیری ماشین را غیرقابل استفاده کند و باعث شود پیش‌بینی‌های نادرست، تصادفی یا بی‌معنی انجام دهد و بدین ترتیب دسترسی آن را برای وظایف مورد نظر به خطر بیندازد.
- **تخلیه منابع:**
  - تلاش‌های مکرر برای آموزش مجدد به منظور کاهش تأثیر حمله می‌تواند منجر به تخلیه منابع شود و زمان، قدرت محاسباتی و تلاش انسانی را مصرف کند بدون اینکه لزوماً عملکرد کامل مدل را بازگرداند.

- **استفاده از آسیب‌پذیری‌های آموزشی:**
  - این حمله از آسیب‌پذیری‌های ذاتی در فرآیند آموزش بهره‌برداری می‌کند، به‌ویژه وابستگی مدل به کیفیت و یکپارچگی داده‌های آموزشی برای یادگیری الگوهای دقیق.

۳-۱۰-۲ دارایی‌های هدف

در حملات مسمومیت دسترس‌پذیری، مهاجمان به دارایی‌های کلیدی در زنجیره یادگیری ماشین هدف قرار می‌دهند تا عملکرد و قابلیت اعتماد مدل‌ها را به خطر بیندازند. این دارایی‌ها برای عملکرد صحیح و پیاده‌سازی سامانه‌های یادگیری ماشین حیاتی هستند و شامل موارد زیر می‌شوند:

- **داده‌های آموزشی:**
  - هدف اصلی مجموعه داده‌ای است که برای آموزش مدل استفاده می‌شود. با معرفی نویز، برچسب‌های نادرست یا ویژگی‌های غیرمربوط، مهاجمان می‌توانند داده‌ها را خراب کرده و باعث شوند مدل الگوهای نادرست را یاد بگیرد. مهاجمان ممکن است از آسیب‌پذیری‌ها در فرآیندهای جمع‌آوری داده یا سامانه‌هایی که داده‌ها در آن‌ها ذخیره می‌شوند، سوءاستفاده کنند تا نمونه‌های مسموم را وارد کنند.
- **فرآیند آموزش مدل:**
  - با خراب کردن داده‌های ورودی در فرآیند آموزش، مهاجمان به طور غیرمستقیم پارامترها و وزن‌های مدل را هدف قرار می‌دهند که منجر به پیکربندی‌های نامطلوب و کاهش عملکرد می‌شود. وجود داده‌های مسموم می‌تواند فرآیند انتخاب یا استخراج ویژگی را گمراه کند و باعث شود مدل بر روی ویژگی‌های غیرمربوط یا گمراه‌کننده تمرکز کند.
- **زنجیره‌های داده و پیش‌پردازش:**
  - مهاجمان ممکن است مراحل تبدیل یا پیش‌پردازش داده‌ها را هدف قرار دهند تا داده‌های معیوب یا مغرضانه را معرفی کرده و اطمینان حاصل کنند که این اثر به کل مجموعه داده و عملکرد مدل‌های بعدی منتقل می‌شود. حملات مسمومیت می‌توانند بر فرآیندهای تقویت داده تأثیر بگذارند و اطمینان حاصل کنند که الگوها یا ناهنجاری‌های مشکل‌ساز در مجموعه‌های تقویت‌شده گنجانده می‌شوند.
- **مدل‌های یادگیری ماشین:**
  - به طور غیرمستقیم، مهاجمان استحکام معماری مدل را هدف قرار می‌دهند و توانایی آن را برای مدیریت تغییرات و داده‌های خراب در مقیاس بزرگ آزمایش می‌کنند. چارچوب‌های نرم‌افزاری و کتابخانه‌های مورد استفاده برای آموزش مدل‌ها می‌توانند هدف آسیب‌پذیری‌هایی

قرار گیرند که اجازه وارد کردن داده‌های مسموم را در طول آموزش می‌دهند.

- **مراحل اعتبارسنجی و آزمایش:**
  - اگر فرآیندهای اعتبارسنجی ضعیف یا مختل شده باشند، ممکن است نتوانند کاهش عملکرد مدل ناشی از داده‌های مسموم را شناسایی کنند و این منجر به پیاده‌سازی مدل‌های معیوب می‌شود. مهاجمان ممکن است به نقاط کور خاص در پروتکل‌های آزمایش هدف قرار دهند تا اطمینان حاصل کنند که مدل مسموم بدون شناسایی از چک‌های معمول عبور می‌کند.
- **محیط پیاده‌سازی:**
  - سامانه‌هایی که داده‌ها را به مدل‌های پیاده‌سازی‌شده تغذیه می‌کنند می‌توانند هدف قرار گیرند تا به تخریب عملکرد پس از پیاده‌سازی ادامه دهند و سناریوهایی مشابه با مسمومیت در زمان آموزش را شبیه‌سازی کنند.
  - با هدف قرار دادن این دارایی‌ها، مهاجمان می‌توانند به طور مؤثر دسترس‌پذیری و قابلیت اعتماد مدل‌های یادگیری ماشین را به خطر بیندازند که منجر به اختلال در خدمات و کاهش عملکرد سامانه می‌شود. ایجاد دفاع‌های قوی و سامانه‌های نظارتی برای این دارایی‌ها برای کاهش تأثیرات حملات مسمومیت دسترس‌پذیری حیاتی است.

۳-۱۰-۳ آسیب‌پذیری‌ها

حملات مسمومیت دسترس‌پذیری از آسیب‌پذیری‌های خاصی در سامانه‌های یادگیری ماشین سوءاستفاده می‌کنند تا عملکرد و دسترسی مدل‌ها را کاهش دهند. درک این آسیب‌پذیری‌ها برای توسعه راهبردهای کاهش این حملات ضروری است. در اینجا برخی از آسیب‌پذیری‌های کلیدی که می‌توانند هدف قرار گیرند، آمده است:

- **کیفیت و یکپارچگی داده: ب**
  - بررسی‌های ناکافی یا عدم وجود آن‌ها برای کیفیت داده، اجازه می‌دهد داده‌های خراب یا دستکاری‌شده خصمانه به مجموعه آموزشی وارد شوند. وابستگی به داده‌های ناشناخته یا منابع احتمالی مختل‌شده، خطر معرفی داده‌های مسموم به سامانه را افزایش می‌دهد.
- **ضعف‌های فرآیند آموزش:**
  - مدل‌هایی که به نویز و نقاط پرت حساس هستند، در مواجهه با داده‌های مسموم در طول آموزش، بیشتر در معرض کاهش عملکرد قرار می‌گیرند. سامانه‌هایی که به داده‌های آموزشی بیش‌برازش می‌کنند، ممکن است به راحتی توسط حتی یک درصد کوچک از داده‌های مسموم گمراه شوند، زیرا مدل ممکن است همبستگی‌های نادرست را یاد بگیرد.

- **عدم استحکام و آزمایش کافی:**
  - بدون آزمایش یا اعتبارسنجی خصمانه دقیق که سناریوهای حمله احتمالی را شامل شود، مدل‌ها ممکن است بدون درک آسیب‌پذیری‌های خود نسبت به داده‌های مسموم پیاده‌سازی شوند. هم‌چنین عدم وجود سازوکارهای شناسایی ناهنجاری قوی برای شناسایی الگوها یا نقاط داده مشکوک در طول آموزش، آسیب‌پذیری را افزایش می‌دهد.
- **نظارت و ثبت نام ناکافی:**
  - نظارت ناکافی بر ورودی داده و زنجیره‌های تبدیل می‌تواند اجازه دهد داده‌های مسموم وارد شده و بدون شناسایی بر آموزش مدل تأثیر بگذارند. علاوه بر این بدون حسابرسی و ثبت دقیق دسترسی و تغییرات داده، شناسایی و ردیابی منبع مسمومیت داده‌ها دشوار می‌شود.
- **پیچیدگی مدل:**
  - در حالی که مدل‌های پیچیده می‌توانند الگوهای پیچیده را ثبت کنند، اما ممکن است به طور ناخواسته الگوهای نادرست را از داده‌های مسموم یاد بگیرند اگر به درستی منظم نشوند. هم‌چنین مدل‌هایی که به استخراج ویژگی‌های قوی توجه نمی‌کنند، بیشتر در معرض گمراهی توسط ورودی‌های مسموم هستند.
- **سازوکارهای پاسخ و به‌روزرسانی:**
  - اقدامات کند یا واکنشی در پاسخ به کاهش عملکرد می‌تواند تأثیر حملات دسترس‌پذیری را تشدید کند. به‌روزرسانی یا آموزش مجدد نادر با مجموعه‌های داده قدیمی تأثیرات داده‌های مسموم را تشدید می‌کند، زیرا مدل ممکن است به کار با اطلاعات قدیمی و خراب ادامه دهد.

۳-۱۰-۴ سناریو تهدید

در یک حمله مسمومیت دسترس‌پذیری، یک مهاجم قصد دارد اثر بخشی مدل یادگیری ماشین را با وارد کردن داده‌های مسموم به مجموعه آموزشی به خطر بیندازد. این حمله هدفش کاهش قابل توجه دقت و قابلیت اعتماد مدل است که بر عملیات و توانایی‌های تصمیم‌گیری یک سازمان تأثیر می‌گذارد.

- **دسترسی و آماده‌سازی:**
  - شناسایی هدف: مهاجم یک سامانه یادگیری ماشین مورد نظر را شناسایی می‌کند، به‌ویژه سامانه‌هایی که منابع داده‌ای در دسترس یا در معرض دید دارند.
  - جمع‌آوری داده: نمونه‌های داده‌ای مشابه با آنچه سامانه استفاده می‌کند جمع‌آوری می‌شود تا الگوها و ساختارهای معمول را درک کند.

- **تولید داده‌های مسموم:**
  - ایجاد ورودی‌های خصمانه: نمونه‌های داده را ایجاد یا تغییر می‌دهند و نویز، اطلاعات نامربوط یا برچسب‌های نادرست را وارد می‌کنند تا نمونه‌های مسموم تولید کنند.
  - تعادل بین پنهان‌سازی و تأثیر: اطمینان حاصل می‌شود که نمونه‌های مخرب با داده‌های مشروع ترکیب شوند تا از شناسایی فرار کنند و در عین حال تأثیر کافی بر یادگیری داشته باشند.
- **تزریق داده:**
  - وارد کردن به زنجیره‌های داده: از دسترسی مستقیم (مانند از طریق جریان‌های داده مختل‌شده) یا روش‌های غیرمستقیم (مانند آلوده کردن مجموعه‌های داده منبع باز) برای وارد کردن داده‌های مسموم به خط لوله آموزشی استفاده می‌شود.
  - سوءاستفاده از ضعف‌ها: هدف قرار دادن سامانه‌هایی که در اعتبارسنجی داده‌ها آسیب‌پذیری دارند یا به داده‌های شخص ثالث وابسته‌اند.
- **آموزش مدل:**
  - پردازش نمونه‌های مسموم: هنگامی که مدل بر روی مجموعه داده ترکیبی آموزش می‌بیند، الگوهای نادرست را یاد می‌گیرد که عملکرد کلی آن را کاهش می‌دهد.
  - مشاهده تأثیر آموزش: اگر ممکن باشد، تغییرات در طول آموزش را تحت نظر قرار دهید تا حمله را برای بیشترین اختلال بهینه کنید.
- **پیاده‌سازی و سوءاستفاده:**
  - کاهش عملکرد: پس از پیاده‌سازی، مدل دقت و قابلیت اعتماد کمتری نشان می‌دهد و نمی‌تواند به طور مداوم نتایج مورد انتظار را ارائه دهد.
  - اختلال عملیاتی: عملکرد مختل شده مدل، عملیات را مختل کرده و بر فرآیندهای تصمیم‌گیری مرتبط با سامانه هوش مصنوعی تأثیر می‌گذارد.

۳-۱۰-۵ نشانه‌های احتمال نفوذ

نشانه‌های احتمال نفوذ برای حملات مسمومیت دسترس‌پذیری، نشانه‌ها یا شواهدی هستند که نشان می‌دهند عملکرد و دسترسی یک مدل یادگیری ماشین به‌طور عمدی کاهش یافته است. در اینجا برخی از این نشانه‌های بالقوه که باید مراقب آن‌ها بود، آورده شده است:

- **ناهنجاری‌های عملکرد:**
  - کاهش قابل توجه در دقت مدل یا افزایش نرخ خطاها بدون توضیح واضح، به‌ویژه پس از

یک جلسه آموزشی اخیر، می‌تواند نشانه‌ای از مسمومیت داده باشد. ناهنجاری‌ها در ثبات پیش‌بینی، جایی که ورودی‌های مشابه خروجی‌های بسیار متفاوتی تولید می‌کنند، ممکن است به فرآیند آموزشی خراب اشاره کند.

- **الگوهای داده غیرعادی:**
  - شناسایی الگوهای غیرمعمول، نقاط پرت یا واریانس بالاتر از حد معمول در داده‌های آموزشی می‌تواند نشان‌دهنده وجود نمونه‌های مسموم باشد. تغییرات قابل توجه و غیرمنتظره در توزیع یا ویژگی‌های داده ممکن است نشانه‌ای خطرناک باشد.
- **اختلالات عملیاتی:**
  - افزایش نرخ خطا یا شکایات کاربران مربوط به خروجی‌های غیرمنتظره یا نادرست مدل می‌تواند نشانه‌ای از وجود مشکل باشد. خرابی‌های مکرر یا ناپایداری در سامانه‌هایی که برای تصمیم‌گیری‌های لحظه‌ای به مدل وابسته هستند، ممکن است نشان‌دهنده نفوذ باشد.
- **تحلیل خروجی‌های مدل:**
  - تحلیل نشان می‌دهد که برای ورودی‌های مشابه، تنوع بالایی در خروجی‌ها وجود دارد که قبلاً نتایج ثابتی تولید می‌کردند. برخی از کلاس‌ها یا خروجی‌ها به‌طور مداوم بیشتر از حد معمول اشتباه دسته‌بندی می‌شوند که ممکن است به داده‌های مسموم هدف‌گذاری شده آن دسته‌ها اشاره کند.
- **هشدارهای نظارت و ثبت‌نام:**
  - ناهنجاری‌های شناسایی شده در لاگ‌ها مربوط به دسترسی داده، تغییرات یا خط لوله آموزشی می‌تواند نشان‌دهنده احتمال دستکاری باشد. هشدارها از سامانه‌های نظارتی که برای شناسایی الگوها یا حجم‌های غیرمعمول جریان داده طراحی شده‌اند، به‌ویژه در طول مرحله آموزشی.
- **مقایسه با معیارها:**
  - ارزیابی‌های معیار یا تست A/B انحرافات قابل توجهی از استانداردهای عملکرد مورد انتظار یا نُرم‌های تاریخی را نشان می‌دهد.
- **اختلافات در ردپای حسابرسی:**
  - ناهنجاری‌ها در فرآیندهای مدیریت داده یا نشانه‌گذاری/پرچم‌گذاری غیرقابل توضیح داده‌ها در طول حسابرسی ممکن است نشان‌دهنده درج داده‌های مختل‌شده باشد. دسترسی

غیرمجاز به داده‌ها یا مدل‌ها، به‌ویژه در اطراف زمان‌های آموزشی، ممکن است نشانه‌ای از حمله باشد.

۳-۱۰-۶ تأثیرات

حملات مسمومیت دسترس‌پذیری می‌توانند تأثیرات گسترده‌ای بر روی سامانه‌های یادگیری ماشین و زمینه‌های وسیع‌تری که در آن‌ها استفاده می‌شوند، داشته باشند. این تأثیرات نه تنها بر روی عملکرد فنی مدل‌ها، بلکه بر جنبه‌های عملیاتی، مالی و شهرت یک سازمان نیز تأثیر می‌گذارد. در اینجا برخی از تأثیرات کلیدی آورده شده است:

- **کاهش عملکرد مدل:**
  - تأثیر فنی اصلی، کاهش قابل توجه دقت مدل است که منجر به پیش‌بینی‌ها و خروجی‌های غیرقابل اعتماد می‌شود. مدل‌های تحت تأثیر ممکن است نرخ خطای بالاتری نشان دهند که آن‌ها را برای اهداف تصمیم‌گیری غیرموثر می‌کند.
- **اختلالات عملیاتی:**
  - سازمان‌ها ممکن است با اختلالاتی در خدمات وابسته به مدل‌های تحت تأثیر مواجه شوند که نیاز به عیب‌یابی فوری و احتمالاً زمان از دست رفته دارد. گردش کارهای عملیاتی که به پیش‌بینی‌های دقیق و به موقع وابسته هستند ممکن است به دلیل خروجی‌های غیرقابل اعتماد با تأخیر مواجه شوند.
- **مصرف منابع و هزینه‌ها:**
  - اصلاح تأثیر حمله مسمومیت دسترس‌پذیری نیاز به منابع محاسباتی و انسانی قابل توجهی دارد، از جمله آموزش مجدد مدل‌ها و پاک‌سازی داده‌ها. نیاز به اقدامات امنیتی اضافی، تحلیل‌های قانونی و احتمالاً از دست دادن کسب‌وکار می‌تواند منجر به هزینه‌های مالی قابل توجهی شود.
- **آسیب به شهرت:**
  - مشکلات مداوم در عملکرد مدل می‌تواند اعتماد مشتریان و ذینفعان را کاهش دهد و بر شهرت سازمان تأثیر بگذارد. در صنایعی که قابلیت اعتماد و دقت حیاتی هستند، مدل مختل‌شده می‌تواند تصویر برند و موقعیت بازار را آسیب بزند.
- **تضعیف تجربه و رضایت کاربر:**
  - کاربران که به خروجی‌های ثابت از سامانه‌های هوش مصنوعی وابسته هستند ممکن است با ناامیدی و نارضایتی مواجه شوند که بر وفاداری و تعامل مشتری تأثیر می‌گذارد. اگر

مدل‌ها به عنوان غیرقابل اعتماد یا مختل‌شده دیده شوند، کاربران ممکن است نگرانی‌هایی در مورد حریم خصوصی و امنیت داده‌های خود پیدا کنند.

- **ریسک‌های نظارتی و انطباق:**
  - اگر خروجی‌های مدل برای رعایت استانداردهای نظارتی یا الزامات انطباق حیاتی باشند، کاهش آن‌ها می‌تواند منجر به نقض‌های نظارتی شود. سازمان‌ها ممکن است با چالش‌های قانونی مواجه شوند اگر مشکلات دسترس‌پذیری منجر به نتایج منفی قابل توجهی برای کاربران شوند یا اگر نتوانند الزامات قراردادی را برآورده کنند.
- **پیامدهای راهبردی بلندمدت:**
  - مشکلات مداوم در دسترس‌پذیری ممکن است منجر به تردید در پذیرش گسترده راه‌حل‌های هوش مصنوعی در سراسر سازمان شود. تلاش‌های لازم برای کاهش حملات و بازگرداندن عملکرد ممکن است منابع را از نوآوری و ابتکارات راهبردی منحرف کند.

۳-۱۰-۷ راه‌های مقابله

کاهش حملات مسمومیت دسترس‌پذیری نیازمند یک رویکرد چندوجهی است که بر تقویت امنیت و استحکام سامانه‌های یادگیری ماشین در سراسر خط داده، فرآیندهای آموزشی و استقرار متمرکز است. در اینجا راهبردهای کلیدی برای کاهش این‌گونه حملات آورده شده است:

- **یکپارچگی داده و تضمین کیفیت:**
  - منابع داده ایمن و تأیید شوند تا یکپارچگی مجموعه‌های داده آموزشی تضمین شود. از منابع معتبر و تأیید شده استفاده شود و به‌طور منظم منبع داده بررسی شوند. سامانه‌های تشخیص ناهنجاری پیاده‌سازی شوند تا نقاط داده مشکوک یا دورافتاده شناسایی و فیلتر شوند که می‌تواند نشان‌دهنده مسمومیت باشد.
- **تکنیک‌های آموزشی مقاوم:**
  - نمونه‌های خصمانه در فرآیند آموزشی گنجانده شوند تا تاب‌آوری مدل در برابر ورودی‌های تحریف‌شده افزایش یابد. از تکنیک‌ها و الگوریتم‌های آماری مقاوم استفاده شود که بتوانند بین نویز معمولی و داده‌های مشکوک تمایز قائل شوند.
- **پیش‌پردازش و تقویت داده:**
  - روش‌های پاک‌سازی سخت‌گیرانه برای حذف داده‌های ناهنجار و بالقوه مضر قبل از استفاده در آموزش به کار روند. از تکنیک‌های تقویت داده برای ایجاد مجموعه‌های داده آموزشی متنوع استفاده شود که به مدل‌ها کمک کند تا بهتر عمومی‌سازی کنند و کمتر به داده‌های

مسموم حساس باشند.

- **تقویت معماری مدل و الگوریتم:**
  - تکنیک‌های تنظیم پیاده‌سازی شوند تا بیش‌برازش را کاهش داده و توانایی مدل را در مقابله با نویز بهبود دهد. از مدل‌های تجمیعی استفاده شود که پیش‌بینی‌های چندین مدل را ترکیب می‌کنند تا تأثیر هر مدل مختل‌شده‌ای را کاهش دهند.
- **اعتبارسنجی و تست مقاوم:**
  - سناریوهای متنوع خصمانه در پروتکل‌های تست گنجانده شده تا تاب‌آوری مدل در برابر انواع مختلف مسمومیت داده ارزیابی شود. از تکنیک‌های اعتبارسنجی متقابل استفاده شود تا استحکام مدل در زیرمجموعه‌های مختلف داده تضمین شود.
- **نظارت و ثبت‌سازی:**
  - به‌طور مداوم خروجی‌های مدل و ورودی‌های داده در زمان واقعی نظارت شده تا رفتارهای ناهنجار به سرعت شناسایی شود. سوابق جامع از تغییرات داده، الگوهای دسترسی و معیارهای عملکرد مدل حفظ شود تا قابلیت ردیابی و تحلیل‌های جنایی تسهیل شود.
- **پاسخ به حادثه و برنامه‌های بازیابی:**
  - برنامه‌های پاسخ به حادثه توسعه و پیاده‌سازی شود تا به سرعت اثرات حملات شناسایی‌شده کاهش یابد. هم‌چنین بایدسازوکارهایی برای آموزش مجدد سریع یا بازگرداندن مدل‌ها به نسخه‌های پایدار قبلی در صورت شناسایی حمله وجود داشته باشد.
- **یادگیری و سازگاری مداوم:**
  - به‌طور مکرر مدل‌ها به‌روزرسانی و با داده‌های تازه و تأیید شده مجدداً آموزش داده شود تا تأثیر هرگونه مسمومیت شناسایی‌نشده قبلی به حداقل برسد. باید در مورد آخرین تهدیدات و آسیب‌پذیری‌های مربوط به مسمومیت داده در جوامع علمی یادگیری ماشین مطلع بود.

۳-۱۱ گریز یا نمونه‌های متخاصم

گریز[1] در یادگیری ماشین به دسته‌ای از حملات اشاره دارد که در آن‌ها مهاجمان داده‌های ورودی را دستکاری می‌کنند تا مدل‌های یادگیری ماشین را فریب دهند. هدف از این دستکاری این است که مدل پیش‌بینی‌ها یا دسته‌بندی‌های نادرستی انجام دهد بدون اینکه خود مدل تغییر کند. حملات گریز به‌ویژه نگران‌کننده هستند زیرا از آسیب‌پذیری‌ها در فرآیندهای تصمیم‌گیری مدل‌ها بهره‌برداری می‌کنند و معمولاً منجر به پیامدهای قابل توجهی در کاربردهای مختلف مانند امنیت سایبری، مالی و

1. Evasion

بهداشت و درمان می‌شوند.

۳-۱۱-۱ ویژگی‌های کلیدی

حملات گریز در یادگیری ماشین ویژگی‌های کلیدی متعددی دارند که آن‌ها را از سایر انواع حملات متمایز می‌کند. درک این ویژگی‌ها برای توسعه دفاع‌های مؤثر و بهبود استحکام مدل‌های یادگیری ماشین ضروری است. در اینجا ویژگی‌های اصلی حملات گریز آورده شده است:

- **دستکاری داده‌های ورودی:**
  - حملات گریز شامل تغییرات ظریف در داده‌های ورودی، مانند تصاویر، متن یا صدا است تا مدل را فریب دهند در حالی که تغییرات برای ناظران انسانی غیرقابل تشخیص باقی می‌ماند. افزودن نویز کوچک به یک تصویر که باعث می‌شود یک طبقه‌بند شیء را نادرست شناسایی کند.
- **حملات هدفمند در مقابل غیرهدفمند:**
  - در حملات هدفمند مهاجم قصد دارد مدل ورودی را در یک کلاس نادرست خاص طبقه‌بندی کند. در حالی که در حملات غیرهدفمند، هدف فقط این است که مدل پیش‌بینی نادرستی انجام دهد، بدون توجه به خروجی خاص.
- **دانش مدل:**
  - حملات جعبه سفید: مهاجم به طور کامل به معماری مدل، پارامترها و داده‌های آموزشی دسترسی دارد که امکان طراحی دقیق نمونه‌های متخاصم را فراهم می‌کند.
  - حملات جعبه سیاه: مهاجم هیچ‌گونه دانش داخلی از مدل ندارد اما می‌تواند آن را پرسش کند و خروجی‌های آن را مشاهده کند و اغلب به قابلیت انتقال نمونه‌های متخاصم از یک مدل به مدل دیگر تکیه می‌کند.
- **اختلال ورودی:**
  - حملات گریز معمولاً به اختلالات کوچک و به‌دقت محاسبه‌شده متکی هستند که ورودی را تغییر می‌دهند بدون اینکه ساختار کلی آن به‌طور قابل توجهی تغییر کند. برای مثال در دسته‌بندی تصویر، تغییراتی که زیر یک آستانه خاص باشند ممکن است توسط انسان‌ها نادیده گرفته شوند اما هنوز هم مدل را فریب دهند.
- **قابلیت انتقال:**
  - نمونه‌های متخاصم تولید شده برای یک مدل ممکن است همچنین سایر مدل‌ها را فریب دهند، حتی اگر آن‌ها بر اساس معماری‌ها یا مجموعه‌های داده آموزشی متفاوتی باشند. این

ویژگی چالشی برای دفاع در برابر حملات گریز ایجاد می‌کند، زیرا حفاظت از یک مدل ممکن است در برابر مدل‌های دیگر مؤثر نباشد.

- **قابلیت انطباق:**
  - راهبردهای گریز می‌توانند به دفاع‌های به‌کاررفته توسط مدل‌ها انطباق یابند و به‌طور مداوم برای دور زدن تدابیر امنیتی تکامل یابند. برای مثال اگر مدلی برای شناسایی و نادیده گرفتن اختلالات خاص آموزش ببیند، مهاجمان ممکن است راه‌های جدیدی برای تغییر ورودی‌ها پیدا کنند تا به اهداف خود برسند.
- **کاربردپذیری در دنیای واقعی:**
  - حملات گریز می‌توانند در حوزه‌های مختلف، از جمله امنیت سایبری (دور زدن سامانه‌های تشخیص نفوذ)، مالی (دستکاری امتیازدهی اعتباری) و وسایل نقلیه خودران (فریب در شناسایی اشیاء) پیامدهای عملی داشته باشند. به عنوان مثال طراحی علائم ترافیکی متخاصم که برای رانندگان انسانی عادی به نظر می‌رسند اما توسط سامانه درک یک وسیله نقلیه خودران نادرست تفسیر می‌شوند.
- **معیارهای ارزیابی:**
  - اثربخشی حملات گریز معمولاً با استفاده از معیارهایی مانند نرخ موفقیت (چقدر مدل فریب خورده است) و غیرقابل تشخیص بودن (چقدر اختلالات برای انسان‌ها قابل شناسایی هستند) ارزیابی می‌شود. تعادل بین این معیارها برای مهاجمان حیاتی است تا اطمینان حاصل کنند که روش‌های آن‌ها هم مؤثر و هم مخفی هستند.

۳-۱۱-۲ دارایی‌های هدف

در زمینه حملات گریز در یادگیری ماشین، دارایی‌های مختلفی قابل شناسایی هستند. این دارایی‌ها عناصری هستند که مهاجمان قصد دارند آن‌ها را دستکاری یا فریب دهند تا به اهداف خود دست یابند. درک این دارایی‌های هدف برای توسعه دفاع‌های مؤثر در برابر چنین حملاتی ضروری است. در اینجا دارایی‌های اصلی در حملات گریز آورده شده است:

- **مدل‌های یادگیری ماشین:**
  - دارایی اصلی که هدف قرار می‌گیرد، مدل یادگیری ماشین آموزش‌دیده است که پیش‌بینی‌ها یا دسته‌بندی‌ها را بر اساس داده‌های ورودی انجام می‌دهد. برای مثال مهاجمان قصد دارند طبقه‌بندهای تصویر یا مدل‌های تحلیل احساسات را با ارائه ورودی‌های متخاصم فریب دهند.

- **داده‌های ورودی:**
  - داده‌های ورودی ارائه‌شده به مدل، از جمله تصاویر، متن، صدا یا انواع دیگر داده‌ها، یک دارایی حیاتی در حملات گریز هستند. تغییر مقادیر پیکسل در یک تصویر یا تغییر چند کلمه در یک سند متنی به منظور تغییر خروجی مدل نمونه‌ای از این مورد هستند.
- **مرزهای تصمیم‌گیری:**
  - آستانه‌ها یا مرزهایی که مدل برای طبقه‌بندی ورودی‌ها استفاده می‌کند، اغلب هدف قرار می‌گیرند، زیرا مهاجمان سعی دارند این مرزها را با ورودی‌های دستکاری‌شده خود عبور دهند. برای مثال طراحی ورودی‌هایی که نزدیک به مرز تصمیم‌گیری قرار دارند تا از آسیب‌پذیری‌های مدل بهره‌برداری کنند.
- **نمایش‌های ویژگی:**
  - نمایش‌های داخلی ویژگی‌های داده که مدل برای انجام پیش‌بینی‌ها استفاده می‌کند، می‌توانند هدف قرار گیرند تا مدل را گیج کنند. دستکاری ویژگی‌های خاص یک ورودی به منظور تغییر نحوه تفسیر داده‌ها توسط مدل نمونه‌ای از این مورد است.
- **داده‌های آموزشی:**
  - گرچه کمتر در حملات گریز رایج است، برخی از مهاجمان ممکن است سعی کنند داده‌های آموزشی یک مدل را تحت تأثیر قرار دهند تا آسیب‌پذیری‌هایی ایجاد کنند که بعداً قابل بهره‌برداری باشند. برای مثال تزریق داده‌های گمراه‌کننده در مرحله آموزشی به منظور اطمینان از یادگیری ارتباطات نادرست توسط مدل.
- **رابطهای کاربری:**
  - رابطهای که کاربران از طریق آن‌ها با مدل تعامل می‌کنند، می‌توانند هدف قرار گیرند، به‌ویژه در کاربردهایی که ورودی‌های کاربر بر نتایج مدل تأثیر می‌گذارند. برای مثال دستکاری فیلدهای ورودی در یک برنامه وب به منظور ارسال نمونه‌های متخاصم به مدل.
- **محیطهای استقرار:**
  - محیطهایی که مدل‌ها در آن‌ها مستقر می‌شوند، از جمله پیکربندی‌های سخت‌افزاری و نرم‌افزاری، می‌توانند هدف قرار گیرند تا از آسیب‌پذیری‌های خاصی بهره‌برداری شود. به عنوان نمونه حمله به زیرساخت‌های پایه برای تزریق ورودی‌های متخاصم به‌طور مستقیم به خط پردازش مدل.

- **سامانه‌های امنیتی:**
  - در زمینه‌های امنیت سایبری، حملات گریز ممکن است سامانه‌های تشخیص نفوذ (IDS) یا دیگر تدابیر امنیتی که به یادگیری ماشین متکی هستند را هدف قرار دهند. برای مثال طراحی ترافیک شبکه‌ای که از شناسایی یک IDS مبتنی بر یادگیری ماشین فرار کند.

۳-۱۱-۳ آسیب‌پذیری‌ها

- **بیش‌برازش مدل:**
  - زمانی که یک مدل به‌طور استثنایی بر روی داده‌های آموزشی عملکرد خوبی دارد اما قادر به تعمیم بر روی داده‌های جدید نیست، ممکن است نسبت به الگوهای خاص ورودی حساسیت بیش از حدی پیدا کند. مهاجمان می‌توانند از این حساسیت بهره‌برداری کرده و ورودی‌هایی را طراحی کنند که نزدیک به توزیع داده‌های آموزشی قرار دارند و منجر به نادرست طبقه‌بندی شوند.
- **آسیب‌پذیری‌های مرزهای تصمیم‌گیری:**
  - مرزهایی که کلاس‌های مختلف را در فضای ویژگی مدل جدا می‌کنند، می‌توانند اشکال پیچیده‌ای داشته باشند و در نتیجه در برابر دستکاری آسیب‌پذیر باشند. مهاجمان می‌توانند نمونه‌های متخاصم را ایجاد کنند که به‌طور راهبردی نزدیک به این مرزها قرار دارند تا پیش‌بینی‌های نادرست ایجاد کنند.
- **کمبود استحکام:**
  - بسیاری از مدل‌های یادگیری ماشین برای مقابله با اختلالات کوچک در داده‌های ورودی طراحی نشده‌اند و این امر آن‌ها را در برابر نمونه‌های متخاصم آسیب‌پذیر می‌کند. ورودی‌هایی که تنها به‌طور جزئی تغییر یافته‌اند می‌توانند منجر به خروجی‌های به‌طور قابل توجهی متفاوت شوند و به این ترتیب به مهاجمان امکان بهره‌برداری از این کمبود استحکام می‌دهند.
- **عدم اعتبارسنجی کافی ورودی:**
  - بررسی‌های ناکافی بر روی داده‌های ورودی قبل از پردازش می‌تواند به نمونه‌های متخاصم اجازه دهد تا تدابیر امنیتی را دور بزنند. مهاجمان می‌توانند ورودی‌های دستکاری‌شده‌ای را ارسال کنند که سامانه به‌طور مناسب آن‌ها را اعتبارسنجی نمی‌کند و این منجر به گریز موفقیت‌آمیز می‌شود.

- **قابلیت انتقال نمونه‌های متخاصم:**
  - نمونه‌های متخاصم تولید شده برای یک مدل می‌توانند اغلب سایر مدل‌ها را نیز فریب دهند، به‌ویژه اگر آن‌ها معماری‌ها یا داده‌های آموزشی مشابهی داشته باشند. این ویژگی به مهاجمان اجازه می‌دهد تا نمونه‌هایی را بدون نیاز به دانش کامل از مدل هدف طراحی کنند و سطح حمله را افزایش دهند.
- **سازوکارهای دفاعی ناکافی:**
  - بسیاری از دفاع‌های موجود در برابر حملات گریز یا ناکارآمد هستند یا توسط مهاجمان انطباق‌پذیر دور زده می‌شوند. دفاع‌های ضعیف ممکن است حس کاذب امنیت ایجاد کنند و سامانه‌ها را در برابر راهبردهای پیچیده گریز آسیب‌پذیر کنند.
- **آسیب‌پذیری‌های ویژگی:**
  - برخی از ویژگی‌ها در داده‌های ورودی ممکن است در فرآیند تصمیم‌گیری مدل به‌طور خاص تأثیرگذار باشند و تغییر این ویژگی‌ها می‌تواند منجر به نادرست طبقه‌بندی شود. مهاجمان می‌توانند این ویژگی‌های بحرانی را شناسایی و دستکاری کنند تا به اهداف خود به‌طور مؤثر دست یابند.
- **محدودیت‌های ادراکی انسانی:**
  - تغییرات متخاصم معمولاً به‌گونه‌ای طراحی می‌شوند که برای ناظران انسانی غیرقابل تشخیص باقی بمانند و از فاصله بین ادراک انسانی و تفسیر مدل بهره‌برداری کنند. مهاجمان می‌توانند ورودی‌هایی را طراحی کنند که برای انسان‌ها عادی به نظر برسند اما توسط مدل نادرست طبقه‌بندی شوند و این امر شناسایی را دشوار می‌سازد.
- **استقرار مدل‌های ایستا:**
  - استقرار مدل‌های ایستا بدون به‌روزرسانی یا آموزش مجدد منظم می‌تواند منجر به آسیب‌پذیری‌هایی شود زیرا محیط و توزیع ورودی‌ها با گذشت زمان تغییر می‌کند. مدل‌ها ممکن است در طبقه‌بندی ورودی‌ها به‌طور دقیق کمتر مؤثر شوند و این امر آن‌ها را برای گریز آسان‌تر می‌کند.
- **پیچیدگی داده‌های دنیای واقعی:**
  - داده‌های واقعی می‌توانند پر از نویز و نامنظم باشند و این امر فرصت‌هایی را برای مهاجمان ایجاد می‌کند تا از ضعف‌های مدل در مدیریت چنین داده‌هایی بهره‌برداری کنند. مدل‌هایی که بر روی پایگاه‌های داده ایده‌آل آموزش دیده‌اند ممکن است در برابر تنوع‌های واقعی دچار

مشکل شوند و این امر امکان حملات گریز موفق را فراهم می‌کند.

۳-۱۱-۴ سناریو تهدید

در یک حمله گریز مدل یادگیری ماشین، مهاجم به‌دنبال دور زدن یا فریب یک سامانه یادگیری ماشین، مانند یک برنامه امنیتی، سامانه تشخیص تقلب، یا هر فرآیند تصمیم‌گیری خودکار است. مهاجم نمونه‌های خصمانه—ورودی‌هایی که برای انسان‌ها عادی به نظر می‌رسند اما منجر به خروجی‌های نادرست از مدل می‌شوند—را تولید می‌کند و بدین ترتیب به اهداف خود، مانند اجتناب از شناسایی یا فعال کردن اقدامات خاص به‌طور نادرست، دست می‌یابد.

- **شناسایی:**
  - مهاجم مدل هدف را بررسی می‌کند تا معماری، ویژگی‌های ورودی و رفتار آن را درک کند. این ممکن است شامل آزمودن مدل از طریق یک API عمومی برای جمع‌آوری بینش‌هایی درباره چگونگی تمایز آن بین کلاس‌ها باشد.
- **تولید مثال‌های خصمانه:**
  - با استفاده از تکنیک‌هایی مانند روش‌های مبتنی بر گرادیان (به‌عنوان مثال، روش علامت گرادیان سریع، کاهش گرادیان پیش‌بینی‌شده)، مهاجم ورودی‌های مشروع را تغییر می‌دهد تا مثال‌های خصمانه ایجاد کند. این مثال‌ها به‌طور خاص طراحی شده‌اند تا پیش‌بینی‌های مدل را تغییر دهند در حالی که به ورودی اصلی نزدیک باقی می‌مانند.
- **آزمایش و تصحیح:**
  - مهاجم مثال‌های خصمانه را در برابر مدل آزمایش می‌کند تا اطمینان حاصل کند که آن‌ها به‌طور مداوم منجر به طبقه‌بندی نادرست می‌شوند. اگر تلاش‌های اولیه مؤثر نباشد، مهاجم نمونه‌ها را تصحیح می‌کند تا موفقیت در فرار را بهبود بخشد.
- **استقرار مثال‌های خصمانه:**
  - پس از تصحیح، ورودی‌های خصمانه در سناریوهای واقعی (به‌عنوان مثال، با دور زدن یک سامانه امنیتی یا دستکاری خروجی‌ها به‌گونه‌ای که به نفع مهاجم باشد) به کار گرفته می‌شوند.

۳-۱۱-۵ نشانه‌های احتمال نفوذ

- **الگوهای ورودی غیرمعمول:**
  - تغییرات ناگهانی در توزیع داده‌های ورودی، مانند افزایش ناگهانی در ویژگی‌های خاص یا

ناهنجاری‌ها در الگوهای داده. برای مثال افزایش قابل توجه در نمونه‌های ورودی که بسیار نزدیک به مرزهای تصمیم‌گیری قرار دارند یا شامل اختلالات جزئی هستند.

- **کاهش عملکرد مدل:**
  - کاهش قابل توجه در دقت یا قابلیت اطمینان پیش‌بینی‌های مدل، به‌ویژه هنگام پردازش داده‌هایی که قبلاً به‌خوبی طبقه‌بندی شده‌اند. برای مثال یک طبقه‌بند که قبلاً دقت بالایی داشت، ناگهان تعداد قابل توجهی از ورودی‌ها را نادرست طبقه‌بندی می‌کند.
- **حجم یا نرخ بالای درخواست‌ها:**
  - تعداد غیرمعمولی از درخواست‌ها یا جستجوها به مدل در یک بازه زمانی کوتاه، که ممکن است نشان‌دهنده تلاش برای شناسایی نقاط ضعف باشد. افزایش ناگهانی در درخواست‌ها که بر روی موارد حاشیه‌ای یا گوشه‌ای داده‌های ورودی تمرکز دارد.
- **الگوهای خروجی غیرمعمول:**
  - خروجی‌های غیرمنتظره یا ناسازگار از مدل که با رفتار عادی عملیاتی هم‌خوانی ندارد. مدل شروع به طبقه‌بندی ورودی‌های بی‌خطر به عنوان مخرب یا بالعکس می‌کند که نشان‌دهنده احتمال دستکاری است.
- **ابزارهای تولید نمونه‌های متخاصم:**
  - شناسایی ابزارها یا کتابخانه‌های خاصی که برای تولید نمونه‌های متخاصم در محیط استفاده می‌شوند. ثبت‌های ورود که نشان‌دهنده استفاده از کتابخانه‌هایی مانند Foolbox یا Clev-erHans هستند که معمولاً برای ایجاد ورودی‌های متخاصم استفاده می‌شوند.
- **افزایش نرخ نادرست‌طبقه‌بندی:**
  - افزایش فراوانی نادرست‌طبقه‌بندی‌ها، به‌ویژه در ورودی‌هایی که قبلاً به‌درستی طبقه‌بندی شده‌اند. برای مثال افزایش ناگهانی در مثبت‌های کاذب یا منفی‌های کاذب در یک وظیفه طبقه‌بندی.
- **ناهنجاری‌های حلقه بازخورد:**
  - الگوهای غیرمعمول در نحوه استفاده از خروجی‌های مدل برای اطلاع‌رسانی به ورودی‌های بعدی، که نشان‌دهنده احتمال دستکاری متخاصم است. داده‌های ورودی که به نظر می‌رسد بر اساس خروجی‌های قبلی مدل طراحی شده‌اند نه داده‌های مستقل و طبیعی نمونه ای از این مورد هستند.

- **تغییرات در اهمیت ویژگی‌ها:**
  - تغییرات قابل توجه در اهمیت یا وزن اختصاص داده شده به ویژگی‌های خاص در فرآیند تصمیم‌گیری مدل. برای مثال ویژگی‌هایی که قبلاً غیرمهم تلقی می‌شدند ناگهان بر نتایج طبقه‌بندی تأثیر قابل توجهی دارند.
- **ناهنجاری‌های رفتار کاربر:**
  - تغییرات در الگوهای رفتار کاربر که با تغییر عملکرد مدل هم‌خوانی دارد، مانند ارسال ورودی‌های مشابه توسط چندین کاربر. برای مثال گروهی از حساب‌ها ورودی‌های یکسان یا تقریباً یکسانی را ارسال می‌کنند که به‌طور جزئی از موارد استفاده معمولی متفاوت است.
- **ناهنجاری‌های لاگ:**
  - ورودی‌های غیرمعمول در لاگ‌هاس دسترسی، مانند تلاش‌های مکرر برای دسترسی به مناطق حساس یک مدل یا سامانه در یک بازه زمانی کوتاه. برای مثال لاگ‌های که نشان‌دهنده تلاش‌های متعدد برای طبقه‌بندی ورودی‌ها یا دسترسی به APIهای مدل هستند که نشان‌دهنده رفتار کاوش‌گرانه است.

۶-۱۱-۳ تأثیرات

- **خسارات مالی:**
  - حملات گریز می‌توانند به‌طور مستقیم منجر به خسارات مالی برای سازمان‌ها به دلیل تقلب، نقض داده‌ها یا زمان‌های عدم دسترسی شوند. مؤسسات مالی ممکن است از معاملات تقلبی که توسط تاکتیک‌های گریز شناسایی نشده‌اند، متحمل خسارات شوند.
- **آسیب به شهرت:**
  - حملات گریز موفق می‌توانند به شهرت یک سازمان آسیب برسانند و منجر به از دست دادن اعتماد و اطمینان مشتریان شوند. برای مثال شرکتی که در یافتن یک نقض داده ناتوان است، ممکن است با واکنش منفی از طرف مشتریان و ذینفعان مواجه شود و تصویر برندش آسیب ببیند.
- **اختلال در عملیات:**
  - حملات گریز می‌توانند عملیات عادی را مختل کنند و منجر به زمان‌های عدم دسترسی یا کاهش عملکرد خدمات و سامانه‌ها شوند. برای مثال یک وسیله نقلیه خودران که به دلیل ورودی‌های متخاصم، علائم راه را نادرست شناسایی می‌کند، می‌تواند منجر به تصادفات و توقف عملیات شود.

- **عواقب قانونی و نظارتی:**
  - سازمان‌ها ممکن است با اقدامات قانونی یا جریمه‌های نظارتی مواجه شوند اگر نتوانند از داده‌های حساس محافظت کرده یا با استانداردهای صنعتی مطابقت کنند..
- **آسیب به سامانه‌های امنیتی:**
  - حملات گریز که به سامانه‌های امنیتی (مانند سامانه‌های تشخیص نفوذ) هدف قرار می‌گیرند، می‌توانند منجر به نقض‌های موفق و دسترسی غیرمجاز شوند.
- **تأثیر بر یکپارچگی مدل:**
  - حملات گریز می‌توانند یکپارچگی و قابلیت اطمینان مدل‌های یادگیری ماشین را تضعیف کنند و باعث تولید نتایج نادرست شوند. برای مثال یک سامانه تشخیص پزشکی که به دلیل ورودی‌های متخاصم، نتایج منفی کاذب ارائه می‌دهد می‌تواند به مراقبت نادرست از بیماران منجر شود.
- **افزایش هزینه‌های دفاعی:**
  - سازمان‌ها ممکن است با هزینه‌های اضافی برای پیاده‌سازی تدابیر امنیتی و دفاع در برابر حملات گریز مواجه شوند. به عنوان نمونه می‌توان به سرمایه‌گذاری در راه‌حل‌های پیشرفته نظارت، آموزش مجدد مدل‌ها، یا استخدام کارشناسان امنیتی برای تقویت چارچوب‌های تدافعی اشاره کرد.
- **از دست دادن مالکیت معنوی:**
  - حملات گریز ممکن است به رقبا یا بازیگران مخرب اجازه دسترسی به الگوریتم‌ها، مدل‌ها یا داده‌های اختصاصی را بدهد. برای مثال یک حمله متخاصم می‌تواند اطلاعات حساس را از یک مدل استخراج کند و به رقبا اجازه دهد تا آن را کپی یا بهره‌برداری کنند.
- **کاهش اعتماد مشتری:**
  - حوادث مکرر حملات گریز می‌تواند موجب از دست رفتن ایمان مشتریان به توانایی یک سازمان در حفاظت از داده‌های آن‌ها و ارائه خدمات قابل اعتماد شود. کاربران ممکن است از پلتفرم‌هایی که به دلیل تاکتیک‌های گریز، نقض‌های مکرر امنیتی داشته‌اند، کنار بکشند.
- **چالش‌ها در پذیرش هوش مصنوعی:**
  - نگرانی‌ها درباره حملات گریز ممکن است مانع از پذیرش فناوری‌های یادگیری ماشین در بخش‌های حیاتی شود. سازمان‌ها ممکن است به دلیل ترس از دستکاری و خطرات مرتبط، از پیاده‌سازی هوش مصنوعی در حوزه‌های بهداشت و درمان یا مالی خودداری کنند.

۳-۱۱-۷ راه‌های مقابله

برای دفاع در برابر حملات گریز در یادگیری ماشین، سازمان‌ها می‌توانند مجموعه‌ای از راهبردها را پیاده‌سازی کنند. این راهبردها با هدف افزایش استحکام مدل، بهبود اعتبارسنجی ورودی و ایجاد یک محیط عملیاتی امن‌تر طراحی شده‌اند. در اینجا برخی از تدابیر مؤثر آورده شده است:

- **آموزش متخاصم:**
  - وارد کردن نمونه‌های متخاصم به مجموعه داده‌های آموزشی تا مدل یاد بگیرد که ورودی‌های دستکاری شده را شناسایی و در برابر آن‌ها مقاومت کند. به‌روزرسانی مداوم مجموعه آموزشی با نمونه‌های تمیز و متخاصم برای بهبود مقاومت مدل راه حلی برای این موضوع است.
- **اعتبارسنجی و پاک‌سازی ورودی:**
  - پیاده‌سازی تکنیک‌های قوی اعتبارسنجی ورودی برای شناسایی و رد ورودی‌های غیرعادی یا مشکوک قبل از رسیدن به مدل. برای این منظور می‌توان از تکنیک‌هایی مانند بررسی دامنه ورودی، اعتبارسنجی نوع و شناسایی ناهنجاری برای فیلتر کردن ورودی‌های بالقوه مضر استفاده کرد.
- **تکنیک‌های تجمیع مدل:**
  - استفاده از چندین مدل یا طبقه‌بند برای انجام پیش‌بینی‌ها که احتمال نادرست‌طبقه‌بندی به دلیل گریز را کاهش می‌دهد و ترکیب خروجی‌های مدل‌های مختلف و استفاده از رأی‌گیری اکثریت یا میانگین‌گیری برای دستیابی به پیش‌بینی‌های قوی‌تر.
- **فشرده‌سازی ویژگی‌ها:**
  - کاهش پیچیدگی ویژگی‌های ورودی با اعمال تغییراتی که جزئیات غیرضروری را حذف می‌کند و در نتیجه سطح حمله را کاهش می‌دهد. تکنیک‌هایی مانند کاهش عمق رنگ تصاویر یا اعمال همواری فضایی می‌توانند به کاهش ورودی‌های متخاصم کمک کنند.
- **تکنیک‌های منظم‌سازی:**
  - اعمال روش‌های منظم‌سازی در حین آموزش برای تشویق مدل به تعمیم بهتر و کاهش حساسیت به اختلالات جزئی. استفاده از تکنیک‌هایی مانند منظم‌سازی L۱ یا L۲ برای تنبیه مدل‌های بسیار پیچیده که ممکن است به حملات گریز آسیب‌پذیرتر باشند.
- **آزمایش و ارزیابی استحکام:**
  - به‌طور منظم مدل را در برابر تکنیک‌های شناخته شده حمله متخاصم آزمایش کنید تا نقاط

ضعف و زمینه‌های بهبود را شناسایی کنید. انجام آزمایش‌های نفوذ و اعتبارسنجی متخاصم برای ارزیابی عملکرد مدل تحت سناریوهای حمله.

- **آستانه‌گذاری پویا:**
  - پیاده‌سازی آستانه‌های پویا برای تصمیم‌گیری‌های طبقه‌بندی بر اساس عوامل زمینه‌ای به منظور کاهش تأثیر ورودی‌های متخاصم و تنظیم مرزهای تصمیم‌گیری در زمان واقعی بر اساس ویژگی‌های ورودی یا وضعیت سامانه برای افزایش استحکام.
- **نظارت و مدیریت لاگ:**
  - به‌طور مداوم عملکرد مدل و الگوهای ورودی را نظارت کنید تا ناهنجاری‌هایی که ممکن است نشان‌دهنده یک حمله گریز در حال انجام باشد، شناسایی شوند. ایجاد سازوکارهای مدیریت لاگ برای پیگیری داده‌های ورودی، پیش‌بینی‌های مدل و بازخورد برای تجزیه و تحلیل در زمان واقعی به این موضوع کمک می‌کند.
- **آموزش و آگاهی کاربران:**
  - آموزش کاربران و ذینفعان درباره خطرات بالقوه مرتبط با سامانه‌های یادگیری ماشین و ترویج بهترین شیوه‌ها با برگزاری جلسات آموزشی و ارائه منابع برای افزایش آگاهی درباره تهدیدات متخاصم و تشویق به استفاده ایمن.
- **راهبردهای دفاعی مشترک:**
  - شرکت در تبادل اطلاعات و همکاری با سایر سازمان‌ها و محققان برای به‌روز ماندن در مورد تهدیدات و دفاع‌های نوظهور و شرکت در فروم‌های صنعتی، پلتفرم‌های تبادل اطلاعات تهدید و ابتکارات تحقیقاتی مشترک برای تقویت امنیت جمعی.

۳-۱۲ حملات مبتنی بر تولید تقویت‌شده با بازیابی(RAG)[1]

حملات مبتنی بر RAG به راهبردهای خصمانه‌ای اشاره دارد که هدف آن‌ها سامانه‌های مولد تقویت‌شده با بازیابی است. تولید تقویت‌شده با بازیابی یک رویکرد یادگیری ماشین است که سازوکار بازیابی را با یک مدل مولد ترکیب می‌کند تا عملکرد وظایفی مانند پاسخ به سوالات و خلاصه‌سازی اسناد را بهبود بخشد. این سامانه اطلاعات یا اسناد مرتبط را از یک مجموعه بزرگ بازیابی کرده و از آن اطلاعات برای تولید پاسخ‌های دقیق‌تر و با زمینه بهتر استفاده می‌کند. این حملات می‌توانند کارایی، قابلیت اطمینان و امنیت سامانه‌های RAG را کاهش دهند و نیاز به دفاع‌های قوی را برجسته می‌کنند.

1. Retrieval-Augmented Generation (RAG)

۳-۱۲-۱ ویژگی‌های کلیدی

حملات مبتنی بر RAG به سامانه‌های تولید تقویت‌شده با بازیابی حمله می‌کنند و از معماری منحصر به فرد آن‌ها که مؤلفه بازیابی را با یک مدل مولد ترکیب می‌کند، بهره‌برداری می‌کنند. در اینجا برخی از ویژگی‌های کلیدی این حملات آورده شده است:

- **هدف‌گیری نقاط ادغام:**
  - حملات مبتنی بر RAG معمولاً بر نقاط ادغام بین مؤلفه‌های بازیابی و تولید تمرکز دارند و سعی می‌کنند جریان اطلاعات دقیق بین آن‌ها را مختل کنند.
- **مسمومیت داده‌ها:**
  - مهاجمان می‌توانند ورودی‌های مضر یا گمراه‌کننده‌ای به پایگاه داده یا دانش‌نامه‌ای که استفاده می‌شود، وارد کنند و مرحله بازیابی را تحت تأثیر قرار دهند و در نتیجه بر خروجی تولید شده توسط مدل تأثیر بگذارند.
- **دستکاری بازیابی:**
  - این حملات می‌توانند به گونه‌ای طراحی شوند که اطلاعات خاصی بازیابی شوند، با تغییر یا تحریف سازوکارهای بازیابی، و باعث شوند سامانه از زمینه نادرست یا کج‌نمایی برای تولید استفاده کند.
- **ساخت ورودی:**
  - ایجاد پرسش‌ها یا ورودی‌های خاص که احتمالاً باعث می‌شود محتوای بازیابی شده، مدل مولد را به تولید خروجی‌های نادرست، کج‌نما یا مضر سوق دهد.
- **بهره‌برداری از الگوریتم‌های رتبه‌بندی:**
  - این حملات می‌توانند با الگوریتم‌های رتبه‌بندی مورد استفاده در مرحله بازیابی تداخل کنند و به‌طور بالقوه ترتیب اطلاعاتی را که به‌عنوان مرتبط‌ترین یا دقیق‌ترین در نظر گرفته می‌شود، تغییر دهند.
- **گمراه‌سازی زمینه:**
  - به دلیل وابستگی به زمینه، این حملات از ضعف‌های موجود در نحوه تفسیر اطلاعات زمینه‌ای توسط مدل مولد بهره‌برداری می‌کنند. این می‌تواند شامل ارسال سیگنال‌های گمراه‌کننده به مدل درباره اینکه کدام زمینه را اولویت دهد، باشد.
- **وابستگی به یکپارچگی داده‌ها:**
  - سامانه‌های RAG به شدت به یکپارچگی داده‌ها وابسته‌اند؛ بنابراین، هر حمله‌ای که کیفیت

داده‌ها را به خطر بیندازد یا نویز را وارد کند، می‌تواند به‌طور قابل توجهی بر قابلیت اطمینان خروجی‌ها تأثیر بگذارد.

- **دستکاری پاسخ‌ها:**
  - با کنترل داده‌های ورودی یا سازوکار بازیابی، مهاجمان می‌توانند پاسخ‌های تولید شده توسط سامانه را دستکاری کنند و به‌طور بالقوه اطلاعات نادرست یا روایت‌های مضر را در خروجی جاسازی کنند.
- **بهره‌برداری از شکاف‌های آموزشی:**
  - این حملات از شکاف‌های موجود در داده‌های آموزشی مدل بهره‌برداری می‌کنند که می‌تواند آن را به دلیل عدم درنظر گرفتن روش‌های مواجه یا درک نوع خاصی از زمینه، در معرض تغییرات خاص قرار دهد.
- **سطح حمله پیچیده:**
  - به دلیل ماهیت دوگانه سامانه‌های RAG (ترکیب بازیابی با تولید) سطح حمله پیچیده‌تر است و نیاز به رویکردهای چندبعدی برای شناسایی و دفاع در برابر تهدیدات بالقوه دارد.

۳-۱۲-۲ دارایی‌های هدف

حملات مبتنی بر RAG به دارایی‌های مختلفی در سامانه‌های تولید تقویت‌شده بر بازیابی (RAG) هدف می‌زنند که برای عملکرد و کارایی آن‌ها حیاتی هستند. در اینجا دارایی‌های اصلی که مهاجمان ممکن است بر روی آن‌ها تمرکز کنند، آورده شده است:

- **پایگاه داده/مجموعه بازیابی:**
  - مجموعه داده بزرگ یا دانش‌نامه‌ای که مؤلفه بازیابی اطلاعات را از آن می‌کشد. به خطر انداختن این دارایی از طریق سم‌پاشی داده‌ها می‌تواند به‌طور قابل توجهی بر ارتباط و دقت اطلاعات بازیابی شده تأثیر بگذارد.
- **الگوریتم‌های بازیابی:**
  - این الگوریتم‌ها مسئول جستجو و رتبه‌بندی داده‌ها از مجموعه هستند. دستکاری در فرآیند بازیابی می‌تواند به استفاده از اسناد کج‌نما یا نادرست توسط مدل مولد منجر شود.
- **مدل مولد:**
  - مؤلفه‌ای که مسئول تولید خروجی نهایی است. حملات هدف‌گیری این دارایی، ممکن است سعی کنند پاسخ‌های آن را کج‌نما کنند یا از سوگیری‌های موجود در آموزش آن بهره‌برداری کنند.

- **لایه ادغام:**
  - رابط یا سازوکاری که مؤلفه‌های بازیابی و تولید را ادغام می‌کند. اختلال در این لایه می‌تواند به نادرستی در نحوه اعمال زمینه بازیابی شده بر تولید منجر شود.
- **داده‌ها و فرآیندهای آموزشی:**
  - هر دو مؤلفه تولیدی و بازیابی به داده‌های آموزشی به‌خوبی تنظیم شده وابسته‌اند. سم‌پاشی داده‌های آموزشی می‌تواند ضعف‌ها یا سوگیری‌های را در بر داشته باشد که مهاجمان می‌توانند بعداً از آن‌ها بهره‌برداری کنند.
- **پارامترها و ابرپارامترهای مدل:**
  - تنظیم این پارامترها می‌تواند بر کیفیت خروجی تأثیر بگذارد. مهاجمان ممکن است در صورت دسترسی، آن‌ها را دستکاری کنند و رفتار مدل را به روش‌های نامطلوب تغییر دهند.
- **کنترل‌های دسترسی و سازوکارهای احراز هویت:**
  - این موارد از این محافظت می‌کنند که چه کسی می‌تواند سامانه را استعلام یا تغییر دهد. ضعف‌ها در اینجا می‌تواند برای معرفی تغییرات غیرمجاز یا اجرای حملاتی مانند ساخت ورودی به کار رود.
- **رابطهای API :**
  - نقاطی که از طریق آن‌ها سامانه‌های خارجی یا کاربران با سامانه RAG تعامل دارند. API ها می‌توانند به عنوان یک مسیر برای حملات تزریقی یا پرسش‌های بدفرم که به منظور بهره‌برداری از ضعف‌های سامانه طراحی شده‌اند، عمل کنند.
- **زیرساخت‌های نظارت و ثبت لاگ:**
  - سامانه‌هایی که فعالیت‌ها و معیارهای عملکرد را ثبت می‌کنند. حملات به این دارایی‌ها می‌تواند شواهد دستکاری را پنهان کند یا از شناسایی فعالیت‌های مخرب جلوگیری کند.
- **سیاست‌ها و پروتکل‌های امنیتی:**
  - سیاست‌ها و پروتکل‌های جامع که تعریف می‌کنند چگونه سامانه ایمن می‌شود. نقص‌ها یا شکاف‌ها در اینجا می‌تواند برای دسترسی غیرمجاز یا اجرای حملات مورد بهره‌برداری قرار گیرد.

۳-۱۲-۳ آسیب‌پذیری‌ها

حملات مبتنی بر RAG از آسیب‌پذیری‌های خاصی که در چارچوب تولید تقویت‌شده با بازیابی (RAG)

نهفته است، بهره‌برداری می‌کنند. درک این آسیب‌پذیری‌ها برای توسعه دفاع‌های قوی ضروری است. در اینجا برخی از آسیب‌پذیری‌های کلیدی مرتبط با سامانه‌های RAG آورده شده است:

- **یکپارچگی داده:**
  - سامانه‌های RAG به شدت به کیفیت و یکپارچگی داده‌ها در پایگاه‌های داده بازیابی خود وابسته‌اند. داده‌های آلوده یا دستکاری شده می‌توانند به خروجی‌های نادرست یا کج‌نما منجر شوند.
- **سوگیری‌های بازیابی:**
  - الگوریتم‌های استفاده شده برای بازیابی اسناد ممکن است سوگیری‌های ذاتی داشته باشند که می‌تواند برای کج‌نمایی اطلاعات ارائه شده به مدل مولد مورد بهره‌برداری قرار گیرد و بر خروجی‌ها تأثیر بگذارد.
- **بهره‌برداری از الگوریتم‌های رتبه‌بندی:**
  - آسیب‌پذیری‌ها در نحوه رتبه‌بندی اسناد برای ارتباط می‌تواند دستکاری شود و باعث شود سامانه برخی نقاط داده که ممکن است دقیق نباشند، بیش از حد اولویت دهد.
- **آسیب‌پذیری‌های آموزشی مدل:**
  - شکاف‌ها یا سوگیری‌ها در داده‌های آموزشی برای هر دو مؤلفه بازیابی و مولد می‌تواند ضعف‌هایی را معرفی کند که مهاجمان ممکن است برای دستکاری رفتار سامانه از آن‌ها بهره‌برداری کنند.
- **وابستگی به زمینه:**
  - از آنجا که سامانه‌های RAG به شدت به اطلاعات زمینه‌ای وابسته‌اند، هرگونه دستکاری در زمینه می‌تواند تأثیر قابل توجهی بر فرآیند تولید و کیفیت خروجی داشته باشد.
- **ساخت ورودی:**
  - این سامانه‌ها ممکن است در برابر ورودی‌ها یا پرسش‌های خصمانه که به‌طور خاص طراحی شده‌اند تا مؤلفه بازیابی یا تولید را گیج یا گمراه کنند، آسیب‌پذیر باشند.
- **اعتبارسنجی ضعیف ورودی:**
  - اعتبارسنجی ناکافی ورودی‌ها می‌تواند اجازه دهد که پرسش‌ها یا دستورات مضر که به‌منظور بهره‌برداری از آسیب‌پذیری‌های فرآیندهای بازیابی یا مولد طراحی شده‌اند، وارد شوند.
- **ضعف‌های احراز هویت و کنترل دسترسی:**
  - کنترل‌های امنیتی ناکافی می‌تواند منجر به دسترسی غیرمجاز به اجزای سامانه شود و به‌طور

بالقوه به دستکاری داده‌ها یا زیر پا گذاشتن پاسخ‌های مدل منجر شود.

- **آسیب‌پذیری‌های API:**
  - APIهای نمایان، اگر به‌درستی ایمن نشده باشند، می‌توانند از طریق حملات تزریقی یا پردازش نادرست پارامترهای درخواست مورد بهره‌برداری قرار گیرند و بر هر دو مؤلفه بازیابی و تولید تأثیر بگذارند.
- **نظارت و ثبت لاگ ناکافی:**
  - عدم وجود نظارت قوی و شناسایی ناهنجاری‌ها می‌تواند از شناسایی به‌موقع حملات یا دستکاری‌های غیرمجاز در سامانه جلوگیری کند.
- **پیچیدگی الگوریتمی:**
  - پیچیدگی ادغام بازیابی با مدل‌های مولد، خطاها و نقاط کور بالقوه‌ای را در نحوه جریان و تبدیل داده‌ها در سامانه به وجود می‌آورد.
- **محدودیت‌های مقیاس‌پذیری:**
  - با افزایش مقیاس سامانه، اطمینان از امنیت و یکپارچگی مداوم در میان منابع داده و اجزای مدل چالش‌برانگیزتر می‌شود و ممکن است آسیب‌پذیری‌های اضافی را باز کند.

۴-۱۲-۳ سناریو حمله

در سناریوی حمله مبتنی بر RAG، یک مهاجم قصد دارد خروجی یک سامانه تولید تقویت‌شده با بازیابی را دستکاری کند. مهاجم به هدف قرار دادن سامانه RAG با تضعیف یکی یا چند مؤلفه، مانند پایگاه داده بازیابی، سازوکار بازیابی یا مدل مولد می‌پردازد تا پاسخ‌های کج‌نما، نادرست یا گمراه‌کننده تولید کند.

- **شناسایی و دسترسی:**
  - مهاجم اطلاعاتی درباره معماری سامانه RAG جمع‌آوری می‌کند، از جمله ماهیت منابع داده، سازوکار بازیابی و هرگونه API یا رابط عمومی نمایان. آن‌ها نقاط ورودی بالقوه برای حمله، مانند نقاط انتهایی ناامن یا مجوزهای دسترسی را شناسایی می‌کنند.
- **سم‌پاشی داده:**
  - مهاجم داده‌های دستکاری شده یا گمراه‌کننده را به پایگاه داده بازیابی وارد می‌کند. این شامل دسترسی مستقیم به پایگاه داده یا بهره‌برداری از ضعف‌های موجود در فرآیند ورود داده برای معرفی داده‌های آلوده است که می‌تواند فرآیند بازیابی را کج‌نما کند.

- **بهره‌برداری از سازوکارهای بازیابی:**
  - مهاجم ممکن است به الگوریتم‌ها یا فرآیندهای رتبه‌بندی و انتخاب اطلاعات از مجموعه هدف قرار دهد. با درک منطق رتبه‌بندی بازیابی، آن‌ها می‌توانند داده‌های آلوده را به‌گونه‌ای طراحی کنند که اولویت بیشتری داشته باشد یا بیشتر بازیابی شود، و بدین ترتیب بر زمینه ارائه شده به مدل مولد تأثیر بگذارند.
- **ساخت ورودی‌های خصمانه:**
  - مهاجم ورودی‌ها یا پرسش‌های خاصی را طراحی می‌کند تا شانس استفاده مدل مولد از داده‌های آلوده یا دستکاری شده بازیابی شده را به حداکثر برساند و بر خروجی نهایی تأثیر بگذارد.
- **دستکاری خروجی:**
  - زمینه یا ورودی تغییر یافته، مدل مولد را مجبور به تولید خروجی‌های کج‌نما، نادرست یا مضر می‌کند. این می‌تواند شامل جاسازی اطلاعات نادرست، ترویج روایت‌ها، یا کاهش کارایی و قابلیت اطمینان پاسخ‌های سامانه باشد.

۳-۱۲-۵ نشانه‌های احتمال نفوذ

شناسایی حملات مبتنی بر RAG شامل بررسی نشانه‌های خاصی از نقض امنیتی در سامانه است. این نشانه‌ها می‌توانند سیگنال‌هایی از احتمال دستکاری یا فعالیت‌های مخرب چارچوب تولید تقویت‌شده با بازیابی (RAG) باشند. در اینجا برخی از این نشانه‌ها کلیدی که باید مراقب آن‌ها بود، آورده شده است:

- **الگوهای داده غیرعادی:**
  - الگوهای غیرمعمول در داده‌های موجود در مجموعه بازیابی، مانند افزایش ناگهانی در تغییرات داده یا داده‌های جدیدی که با ویژگی‌های ورودی معمولی همخوانی ندارند، می‌تواند نشان‌دهنده تلاش‌های سم‌پاشی باشد.
- **تغییرات ناگهانی در خروجی مدل:**
  - تغییرات ناگهانی در کیفیت یا دقت خروجی‌های مدل مولد، به‌ویژه اگر کج‌نما یا نامربوط شوند، می‌تواند نشان‌دهنده دستکاری در زمینه بازیابی باشد.
- **فعالیت بازیابی غیرعادی:**
  - ناهنجاری‌ها در فرآیند بازیابی، مانند افزایش غیرمنتظره در پرسش‌های خاص یا فراوانی نامتناسب در بازیابی داده‌ها، می‌تواند نشان‌دهنده حمله به سازوکار بازیابی باشد.

- **منابع غیرمعتبر:**
  - وجود ورودی‌های داده از منابع تأیید نشده یا ناشناس در پایگاه داده بازیابی، که ممکن است برای دستکاری خروجی‌ها وارد شده باشند.
- **درخواست‌های API زیاد:**
  - حجم بالای درخواست‌های API پیچیده یا بدفرم، که ممکن است نشان‌دهنده تلاش‌های بهره‌برداری از آسیب‌پذیری‌ها یا ساخت ورودی‌های خصمانه باشد.
- **الگوهای دسترسی نامنظم:**
  - الگوهای غیرعادی در دسترسی یا تغییر داده‌ها یا اجزای سامانه، به‌ویژه توسط حساب‌هایی بدون مجوز مناسب یا خارج از ساعات عملیاتی معمول.
- **تغییرات در رفتار رتبه‌بندی:**
  - تغییرات غیرمنتظره در نحوه رتبه‌بندی نتایج توسط الگوریتم بازیابی، که نشان‌دهنده دستکاری در منطق رتبه‌بندی یا اسناد تحت تأثیر است.
- **لاگ‌ها و هشدارهای خطا:**
  - خطاهای غیرقابل توضیح یا هشدارهای امنیتی مرتبط با مراحل بازیابی داده، پردازش یا تولید مدل، که ممکن است نشان‌دهنده تلاش‌های بهره‌برداری زیرساختی باشد.
- **کاهش اعتماد یا تعامل کاربر:**
  - کاهش ناگهانی در اعتماد یا تعامل کاربر به دلیل خروجی‌های نادرست یا ناخواسته مکرر از سامانه، که ممکن است به دلیل فرآیندهای آسیب‌دیده باشد.
- **بازرسی‌های امنیتی و ناهنجاری‌ها:**
  - یافته‌های حاصل از بازرسی‌های امنیتی که تغییرات پیکربندی، ناهنجاری‌ها یا انحرافات از فرآیندها و پروتکل‌های استاندارد را نشان می‌دهد.

۳-۱۲-۶ تأثیرات

حملات مبتنی بر RAG می‌توانند تأثیرات قابل توجه و گسترده‌ای بر عملکرد فنی سامانه‌های تولید تقویت‌شده با بازیابی (RAG) و جنبه‌های عملیاتی و راهبردی گسترده‌تر سازمان‌های استفاده‌کننده از آن‌ها داشته باشند. در اینجا برخی از تأثیرات کلیدی آورده شده است:

- **کاهش عملکرد سامانه:**
  - تأثیر فنی اصلی، کاهش عملکرد سامانه است. خروجی‌ها ممکن است کج‌نما، نادرست یا

نامربوط شوند، که منجر به کاهش کلی کاربرد و دقت سامانه می‌شود.

- **توزیع اطلاعات نادرست:**
  - با دستکاری در فرآیند بازیابی یا تولید، سامانه ممکن است به‌طور غیرعمدی اطلاعات نادرست یا محتوای کج‌نما را گسترش دهد، که با اهداف مهاجم همخوانی دارد و بر ادراک یا تصمیمات کاربران تأثیر می‌گذارد.
- **از دست رفتن اعتماد کاربر:**
  - قرارگیری مداوم در معرض خروجی‌های نادرست یا نامعتبر می‌تواند اعتماد کاربر به سامانه را کاهش دهد، که برای هر کاربردی که به بینش‌ها یا حمایت‌های مبتنی بر داده وابسته است، حیاتی است.
- **اختلالات عملیاتی:**
  - حملات ممکن است منجر به اختلال در گردش‌کارهایی شوند که به سامانه RAG وابسته‌اند و باعث تأخیرها، ناکارآمدی‌ها یا نیاز به تغییر به فرآیندهای دستی شوند.
- **هزینه‌های مالی:**
  - سازمان‌ها ممکن است هزینه‌های مالی مرتبط با بررسی و کاهش اثرات حمله را متحمل شوند، مانند آموزش مجدد مدل‌ها، تجدید ورودی‌های داده، یا تقویت تدابیر امنیتی.
- **آسیب به اعتبار:**
  - اگر سامانه محتوای مضر یا گمراه‌کننده‌ای تولید کند، می‌تواند به اعتبار سازمان آسیب برساند، به‌ویژه اگر این خروجی‌ها بر کاربران نهایی یا خدمات عمومی تأثیر بگذارد.
- **معضل رقابتی:**
  - اگر رقبا سامانه‌های قوی و قابل اعتمادی را حفظ کنند در حالی که سامانه‌های یک سازمان در معرض خطر قرار دارند، ممکن است منجر به از دست دادن مزیت رقابتی در بازار شود.
- **مسائل قانونی و انطباق:**
  - تولید خروجی‌های مضر یا نادرست ممکن است به چالش‌های قانونی منجر شود، به‌ویژه اگر با مقررات صنعتی مغایرت داشته باشد یا به کاربران آسیب برساند.
- **بهره‌برداری برای حملات بیشتر:**
  - سامانه‌های RAG آسیب‌دیده می‌توانند به‌عنوان مسیرهایی برای حملات سایبری گسترده‌تر عمل کنند و به مهاجمان امکان دهند که از اطلاعات نادرست برای مهندسی اجتماعی یا سایر

فعالیت‌های مخرب استفاده کنند.

- **انحراف منابع:**
  - پرداختن به حمله منابع را از سایر ابتکارات راهبردی منحرف می‌کند، زیرا تیم‌های IT و امنیت بر روی بررسی، کاهش و تقویت تدابیر پیشگیرانه تمرکز می‌کنند.

۳-۱۲-۷ راه‌های مقابله

- کاهش حملات مبتنی بر RAG نیاز به یک رویکرد جامع دارد که هم اجزای جستجو و هم تولید را در سامانه‌های تولید با استفاده از جستجو تأمین کند. در اینجا راهبردهای کلیدی برای کاهش این حملات آورده شده است:

- **اعتبارسنجی داده‌ها و بررسی‌های یکپارچگی:**
  - سازوکارهای اعتبارسنجی قوی پیاده‌سازی شده تا از یکپارچگی و کیفیت داده‌های استفاده شده در پایگاه‌های داده جستجو اطمینان حاصل شود. به‌طور منظم داده‌ها بررسی و پاک‌سازی گردد تا ناهنجاری‌ها یا ورودی‌های بالقوه مخرب حذف شوند.
- **الگوریتم‌های جستجوی امن:**
  - الگوریتم‌های جستجویی توسعه و پیاده‌سازی شود که در برابر دستکاری مقاوم باشند. اطمینان حاصل گردد که این الگوریتم‌ها می‌توانند سوگیری یا تأثیرات خصمانه را در فرآیند رتبه‌بندی و انتخاب شناسایی و کاهش دهند.
- **سامانه‌های تشخیص ناهنجاری:**
  - از تکنیک‌های پیشرفته تشخیص ناهنجاری برای شناسایی الگوهای غیرمعمول در دسترسی به داده‌ها، فراوانی پرسش‌ها و خروجی‌های جستجو استفاده شود. این می‌تواند در تشخیص سریع احتمال مسمومیت داده‌ها یا دستکاری‌های جستجو کمک کند.
- **کنترل‌های دسترسی و احراز هویت:**
  - کنترل‌های دسترسی تقویت شده تا تنها کاربران مجاز بتوانند منابع داده یا پیکربندی‌های سامانه را تغییر دهند. پروتکل‌های احراز هویت قوی پیاده‌سازی گردد تا از دسترسی غیرمجاز یا دستکاری جلوگیری شود.
- **پاک‌سازی و اعتبارسنجی ورودی:**
  - تکنیک‌های سخت‌گیرانه پاک‌سازی و اعتبارسنجی ورودی اعمال شده تا از دستکاری سامانه توسط پرسش‌های خصمانه جلوگیری شود. این شامل فیلتر کردن و بررسی دقیق تمام ورودی‌ها قبل از پردازش است.

- **آموزش و به‌روزرسانی تدریجی:**
  - از یادگیری تدریجی برای به‌روزرسانی مدل‌ها به‌طور تکراری استفاده شود تا نظارت و شناسایی ناهنجاری‌های معرفی شده در طول به‌روزرسانی داده‌ها یا اصلاح مدل‌ها آسان‌تر شود.
- **گزارش‌گیری و نظارت جامع:**
  - گزارش‌گیری دقیق از دسترسی به داده‌ها، الگوهای جستجو و خروجی‌های مدل فعال گردد. نظارت بر این گزارش‌ها می‌تواند نشانه‌های اولیه‌ای از احتمال نفوذ یا فعالیت‌های غیرمعمول را فراهم کند.
- **مراجعات امنیتی منظم:**
  - مراجعات امنیتی منظم از معماری RAG، از جمله اجزای نرم‌افزاری و سخت‌افزاری، انجام شده تا آسیب‌پذیری‌ها را به‌طور پیشگیرانه شناسایی و برطرف گردد.
- **سازوکارهای افزونگی و تغییر مسیر:**
  - افزونگی در منابع و الگوریتم‌های جستجو پیاده‌سازی شده تا تأثیر هر عنصر آسیب‌دیده به حداقل برسد. این امر اطمینان از قابلیت اطمینان و تداوم سامانه را تضمین می‌کند.
- **سامانه‌های بازخورد کاربر:**
  - سازوکارهایی را در نظر گرفته که به کاربران اجازه می‌دهد خروجی‌های نادرست یا مشکوک را علامت‌گذاری کنند. بازخورد کاربران می‌تواند در شناسایی مسائل بالقوه با داده‌ها یا رفتار مدل ارزشمند باشد.
- **تکنیک‌های آموزش خصمانه:**
  - از آموزش خصمانه استفاده شود تا مدل‌ها را در برابر سناریوهای حمله احتمالی در طول فاز آموزش قرار گیرند و مقاومت آنها در برابر حملات دنیای واقعی تقویت شود.
- **راهبرد دفاع لایه‌ای:**
  - یک رویکرد امنیتی چندلایه را اتخاذ کنید که اقدامات امنیتی شبکه، حفاظت از نقطه پایانی و راهبردهای حفاظت از داده‌ها را ادغام کند و پوشش جامعی در برابر وکتورهای حمله بالقوه ارائه دهد.

۳-۱۳ تزریق مستقیم درخواست

تزریق مستقیم درخواست[1] به تکنیکی اشاره دارد که در آن یک مهاجم ورودی یا درخواست داده شده به یک مدل پردازش زبان طبیعی، مانند یک مدل زبانی بزرگ، را دستکاری می‌کند تا رفتار آن را تغییر دهد یا آن را وادار کند تا پاسخ‌هایی تولید کند که توسط طراحان آن در نظر گرفته نشده است. این کار می‌تواند با طراحی عبارات یا فرمان‌های خاصی انجام شود که مدل را فریب می‌دهند تا از دستورالعمل‌های قبلی چشم‌پوشی کند یا اطلاعات حساس را بازیابی کند. هدف از تزریق مستقیم درخواست معمولاً دور زدن محدودیت‌ها یا استخراج خروجی‌های غیرمجاز یا مضر از مدل است.

۳-۱۳-۱ ویژگی‌های کلیدی

تزریق مستقیم درخواست دارای چندین ویژگی کلیدی است که نحوه اجرای آن و تأثیرات بالقوه‌اش را تعریف می‌کند:

- **دستکاری ورودی:**
  - در هسته‌اش، این تکنیک شامل طراحی دقیق درخواست‌های ورودی است که باعث انحراف مدل زبان از رفتار مورد انتظارش می‌شود.
- **حذف یا تغییر زمینه:**
  - این تکنیک تلاش می‌کند تا زمینه‌ای که توسط درخواست‌ها یا دستورالعمل‌های قبلی ایجاد شده را حذف یا تغییر دهد و مدل را وادار کند که به محتوای تزریق‌شده اهمیت بیشتری بدهد.
- **استفاده از رفتار مدل:**
  - این تکنیک از طراحی مدل برای تکمیل خودکار، حدس قصد یا پیروی از دستورالعمل‌های ضمنی بر اساس آخرین ورودی استفاده می‌کند.
- **استخراج اطلاعات حساس:**
  - در برخی موارد، این تکنیک می‌تواند برای استخراج اطلاعاتی استفاده شود که باید محافظت‌شده یا پنهان بماند، با فریب دادن مدل به افشای چنین داده‌هایی.
- **عبور از محدودیت‌ها:**
  - درخواست‌ها به‌گونه‌ای طراحی شده‌اند که محدودیت‌ها یا تدابیر ایمنی که در مدل برای جلوگیری از تولید محتوای مضر یا نامناسب وجود دارد را دور بزنند.

1. Direct Prompt Injection

- **محرک‌های زمینه‌ای:**
  - این تکنیک ممکن است از عبارات یا ساختارهای خاصی استفاده کند که به‌طور مؤثری پاسخ‌های مدل را بر اساس داده‌های آموزشی‌اش تغییر می‌دهند.
- **ساختاردهی عمدی نادرست:**
  - ورودی‌ها ممکن است عمداً به‌صورت نادرست یا به‌شیوه‌های غیرمعمول ساختاربندی شوند تا مدل را گیج کنند و نتیجه تزریق مطلوب را به‌دست آورند.

۲-۱۳-۳ دارایی‌های هدف

تزریق مستقیم درخواست می‌تواند به اهداف مختلفی که با مدل‌های زبانی مرتبط هستند، حمله کند و درک این اهداف برای محافظت در برابر تهدیدات بالقوه بسیار مهم است. در اینجا برخی از دارایی‌های کلیدی که ممکن است مورد هدف قرار گیرند، آمده است:

- **اطلاعات حساس:**
  - مهاجمان ممکن است تلاش کنند تا اطلاعات حساس یا اختصاصی که مدل به آن دسترسی دارد، مانند داده‌های آموزشی که شامل اطلاعات محرمانه یا شخصی است، استخراج کنند.
- **یکپارچگی مدل:**
  - رفتار و خروجی مدل می‌تواند تحت تأثیر قرار گیرد یا خراب شود که این امر موجب ایجاد مشکل در یکپارچگی و قابلیت اعتماد پاسخ‌هایی می‌شود که تولید می‌کند.
- **رعایت قوانین و تدابیر ایمنی:**
  - دور زدن محدودیت‌ها و تدابیر ایمنی می‌تواند منجر به تولید خروجی‌هایی شود که با دستورالعمل‌ها و مقررات مغایرت دارد و ممکن است عواقب قانونی یا شهرتی به دنبال داشته باشد.
- **مالکیت معنوی:**
  - مهاجمان ممکن است به‌دنبال کشف جزئیات مربوط به الگوریتم‌های اختصاصی، روش‌های پردازش داده یا دیگر مالکیت‌های معنوی موجود در چارچوب مدل باشند.
- **امنیت عملیاتی:**
  - مهاجمان ممکن است از تزریق درخواست‌ها برای به‌دست آوردن اطلاعاتی در مورد جنبه‌های عملیاتی یک سامانه بهره‌برداری کنند که ممکن است آسیب‌پذیری‌ها را نمایان کند.

۳-۱۳-۳ آسیب‌پذیری‌ها

تزریق مستقیم درخواست از برخی آسیب‌پذیری‌ها درون مدل‌های زبانی و محیطهای استقرار آن‌ها بهره‌برداری می‌کند. درک این آسیب‌پذیری‌ها می‌تواند به طراحی دفاع‌های بهتر کمک کند. در اینجا برخی از آسیب‌پذیری‌های کلیدی که در این زمینه وجود دارند، آمده است:

- **ضعف‌های مدیریت ورودی:**
  - مدل‌هایی که ورودی‌های درخواست را به‌طور مناسب پاک‌سازی یا اعتبارسنجی نمی‌کنند، به تزریق بیشتری آسیب‌پذیر هستند، چراکه ممکن است ورودی‌های مضر یا دستکاری‌شده را بدون بررسی پردازش کنند.
- **عدم آگاهی از زمینه:**
  - مدل‌هایی که توانایی نگهداری مؤثر زمینه یا تشخیص بین دستورالعمل‌های مورد اعتماد و فرمان‌های تزریقی را ندارند، به‌راحتی فریب می‌خورند.
- **اتکای بیش از حد به داده‌های ورودی:**
  - اگر یک مدل به شدت به آخرین ورودی برای تولید خروجی خود وابسته باشد، به تزریق درخواست‌هایی که می‌تواند تنظیمات یا دستورالعمل‌های قبلی را نادیده بگیرد، آسیب‌پذیرتر می‌شود.
- **عدم وجود استانداردهای ایمنی قوی:**
  - نبود تدابیر اخلاقی و ایمنی قوی در درون مدل برای جلوگیری از تولید محتوای مضر، نامناسب یا غیرمترقبه به آسیب‌پذیری آن کمک می‌کند.
- **ابهام در پردازش زبان:**
  - مدل‌های زبانی ممکن است ورودی‌های مبهم یا به‌خوبی ساختاربندی شده را به عنوان دستورالعمل‌های مشروع تفسیر کنند، که این امر به حملات از طریق عبارات خلاقانه یا نحو مبهم کمک می‌کند.
- **تحریک کننده‌های پنهان یا غیرمترقبه:**
  - مدل‌ها ممکن است دارای محرک‌های غیر مستند یا پیش‌بینی‌نشده در سازوکارهای پردازش داده‌های خود باشند که می‌توانند از طریق تزریق درخواست به‌کار گرفته شوند.
- **استفاده از حلقه‌های بازخورد:**
  - در محیطهای تعاملی، یک مهاجم ممکن است بر اساس پاسخ‌های مدل، درخواست‌های خود را به‌طور تدریجی بهبود بخشد تا نتایج تزریق مؤثرتری به‌دست آورد.

۳-۱۳-۴ سناریو تهدید

یک مهاجم به یک دستیار مجازی یا چت بات مبتنی بر هوش مصنوعی مولد که برای ارائه توصیه‌های شخصی و یا کاری استفاده می‌شود، حمله می‌کند. هدف این است که دستیار مجازی را به گونه‌ای دستکاری کند که اطلاعات حساس را افشا کند یا اقداماتی غیرمجاز را انجام دهد، مانند تغییر تنظیمات حساب یا دسترسی به داده‌های محدودشده.

- **جاسوسی:**
  - مهاجم تعاملات و درخواست‌های مرسوم چت بات را مطالعه می‌کند تا بفهمد چگونه ورودی‌ها را پردازش می‌کند و به سؤالات کاربران پاسخ می‌دهد.
- **طراحی درخواست‌های مخرب:**
  - مهاجم درخواست‌هایی طراحی می‌کند که به‌منظور گیج کردن زمینه هوش مصنوعی یا شبیه‌سازی یک دستور مدیریتی باشد. این ممکن است شامل جاسازی دستورات پنهان در سؤالات ظاهراً بی‌ضرر باشد.
- **تلاش برای تزریق:**
  - با استفاده از آزمایش و خطا، مهاجم درخواست‌های طراحی‌شده را به چت بات وارد می‌کند و رویکرد خود را بر اساس پاسخ‌های آن اصلاح می‌کند تا به‌تدریج به نتیجه دلخواه برسد.
- **استفاده:**
  - پس از موفقیت در دستکاری چت بات، مهاجم اطلاعات حساس (مانند داده‌های کاربری یا اطلاعات مالی) را استخراج می‌کند یا دستورات غیرمجاز (مانند بازنشانی رمزهای عبور یا تغییر مجوزهای حساب) صادر می‌کند.

۳-۱۳-۵ نشانه‌های احتمال نفوذ

نشانه‌های احتمال نفوذ برای تزریق مستقیم درخواست می‌توانند به شناسایی زمانی که یک مدل زبانی یا سامانه هوش مصنوعی به‌طرز بالقوه‌ای دستکاری شده است، کمک کنند. در اینجا برخی از شاخص‌های کلیدی که باید به آن‌ها توجه داشت، آورده شده است:

- **خروجی‌های غیرعادی مدل:**
  - سامانه پاسخ‌هایی غیرمنتظره، نامناسب یا بی‌ربط به زمینه تولید می‌کند که از رفتار معمول آن انحراف دارد.
- **تنظیم مجدد مکرر زمینه:**
  - وقوع‌های مکرر که در آن مدل به‌نظر می‌رسد که زمینه مکالمه قبلی را فراموش یا نادیده

می‌گیرد، که ممکن است نشانه‌ای از تلاش تزریق باشد.

- **دسترسی غیرمنتظره به داده‌ها:**
  - بازیابی و نمایش اطلاعات حساس یا محدود که نباید در زمینه کاربری کنونی قابل دسترسی باشد.
- **الگوهای ورودی نامنظم:**
  - الگوهای ورودی غیرعادی یا بدفرم، مانند دستورات جاسازی‌شده در پرسش‌ها یا نحو غیرمعمول، که ممکن است نشانه‌هایی از تلاش‌های دستکاری باشد.
- **افزایش نرخ خطا:**
  - افزایش ناگهانی در خطاها یا تلاش‌های ناموفق مدل برای پردازش ورودی‌ها به‌طور مؤثر، که ممکن است به دلیل تلاش‌های مکرر تزریق باشد که قابلیت پردازش آن را مختل کرده است.
- **حجم بالای تعاملات:**
  - افزایش غیرعادی در تعداد تعاملات یا پرسش‌ها، به‌ویژه آن‌هایی که نیاز به مجوزهای بالای سامانه دارند، می‌تواند نشانه‌ای از تلاش‌های نفوذی یا خودکار باشد.
- **ناهنجاری در لاگ‌های دسترسی:**
  - الگوهای دسترسی غیرمعمول در لاگ‌های سامانه، مانند دسترسی به عملکردهای مدیریتی یا عناصر داده بدون مجوز واضح کاربر.
- **تغییرات در رفتار مدل:**
  - تغییرات ناگهانی در نحوه رفتار یا پاسخ‌های مدل، که ممکن است نشانه‌ای از این باشد که به‌طرز موفقیت‌آمیزی به تغییر عملکردهای عادی‌اش وادار شده است.
  - نظارت بر این نشانه‌ها می‌تواند بخشی از یک راهبرد امنیتی وسیع‌تر برای شناسایی و پاسخ سریع به حملات تزریق درخواست باشد.

۶-۱۳-۳ تأثیرات

تزریق مستقیم درخواست می‌تواند تأثیرات قابل‌توجهی بر سامانه‌های استفاده‌کننده از مدل‌های زبان داشته باشد که عواقبی را در جنبه‌های فنی و تجاری به همراه دارد. در اینجا برخی از تأثیرات کلیدی این حملات آمده است:

- **نقض داده و تخلفات حریم خصوصی:**
  - دسترسی غیرمجاز به اطلاعات حساس یا محرمانه می‌تواند منجر به نقض حریم خصوصی

شود و اطلاعات شخصی، اختصاصی یا مالی که باید محافظت شوند، افشا گردد.

- **کاهش یکپارچگی سامانه:**
  - یکپارچگی سامانه هوش مصنوعی ممکن است به خطر بیفتد و به تولید خروجی‌های تحریف‌شده و فرآیندهای تصمیم‌گیری غیرقابل اعتماد منجر شود که اعتماد به سامانه را تضعیف می‌کند.
- **کاهش اعتماد و آسیب به شهرت:**
  - هنگامی که کاربران بدانند که هوش مصنوعی قادر به حفاظت از داده‌هایشان نیست یا پاسخ‌های قابل اعتمادی نمی‌دهد، اعتماد به سازمان یا خدمات به شدت آسیب می‌بیند که بر وفاداری مشتری و شهرت برند تأثیر منفی دارد.
- **اختلال در عملیات:**
  - دستکاری سامانه می‌تواند منجر به بروز مشکلات عملیاتی، مانند اجرای دستورات نادرست یا مضر، فساد داده‌ها یا عدم ارائه خدمات گردد.
- **مسائل قانونی و انطباق:**
  - افشای اطلاعات حساس می‌تواند به عدم انطباق با مقررات حفاظت از داده‌ها منجر شود و عواقب قانونی، از جمله جریمه‌ها و الزام به افشای نقض اطلاعات، به دنبال داشته باشد.
- **زیان مالی:**
  - تأثیرات مالی مستقیم ممکن است شامل هزینه‌های مربوط به جبران خسارت، مانند بازرسی‌های سامانه، تقویت تدابیر امنیتی، جریمه‌های قانونی و کاهش کسب‌وکار به‌دلیل کاهش اعتماد مشتری باشد.
- **انتشار اطلاعات نادرست یا محتوای مضر:**
  - تزریق درخواست‌های موفق می‌تواند تولید و توزیع اطلاعات نادرست یا محتوای مضر را ممکن سازد که می‌تواند بر یکپارچگی اطلاعات تأثیر بگذارد و در نهایت به آسیب‌های اجتماعی منجر شود.
- **افزایش تهدیدات امنیتی گسترده‌تر:**
  - سوءاستفاده از طریق تزریق درخواست ممکن است به‌عنوان یک دروازه برای سایر تهدیدات سایبری عمل کند و سطح حمله را افزایش داده و سامانه را در معرض آسیب‌پذیری‌های بیشتری قرار دهد.

۳-۱۳-۷ راه‌های مقابله

کاهش ریسک‌های مرتبط با تزریق مستقیم درخواست نیاز به یک رویکرد چند وجهی دارد که ترکیبی از تدابیر تکنیکی، رویه‌ای و سازمانی است. در اینجا برخی از روش‌های مؤثر برای مقابله آمده است:

- **اعتبارسنجی و پاک‌سازی ورودی:**
  - تکنیک‌های سخت‌گیرانه اعتبارسنجی ورودی پیاده‌سازی شده تا ورودی‌های پتانسیل مضر یا بدفرم قبل از رسیدن به مدل زبان فیلتر شوند.
- **مدیریت زمینه:**
  - توانایی مدل برای نگهداری و تأیید زمینه در تمام تعاملات تقویت شوند تا از سردرگمی زمینه جلوگیری شود و استمرار جریان مکالمه اصلی تضمین گردد.
- **کنترل دسترسی و مجوزها:**
  - دسترسی به اطلاعات حساس و دستورات مدیریتی بر اساس نقش‌ها و مجوزهای کاربران محدود شده تا در صورت تلاش‌های تزریق، میزان در معرض قرار گرفتن کاهش یابد.
- **تشخیص ناهنجاری:**
  - از سامانه‌های تشخیص ناهنجاری برای شناسایی الگوهای ورودی غیرمعمول یا رفتار مدل که ممکن است نشانه‌ای از تلاش تزریق درخواست باشد، استفاده شود.
- **اقدامات امنیتی چند لایه:**
  - چندین لایه امنیتی، مانند دیوارهای آتش، سامانه‌های تشخیص نفوذ و رمزنگاری برای محافظت از داده‌ها و محدود کردن دسترسی غیرمجاز مستقر شود.
- **بازرسی‌ها و آزمایش‌های منظم:**
  - بازرسی‌های امنیتی منظم و آزمایش‌های نفوذ انجام شده تا آسیب‌پذیری‌های موجود در سامانه هوش مصنوعی شناسایی و به‌طور پیشگیرانه به آن‌ها رسیدگی گردد.
- **آموزش و آگاهی کاربران:**
  - کاربران و ذینفعان را درباره خطرات و نشانه‌های تزریق درخواست آموزش داده و آن‌ها برای تعامل مسئولانه با سامانه‌های هوش مصنوعی تربیت شوند.
- **لاگ‌گذاری و نظارت قوی:**
  - لاگ‌گذاری جامعی از تعاملات پیاده‌سازی شده و به‌طور مستمر بر فعالیت‌های غیرمعمول یا ناهنجاری‌های دسترسی نظارت شود تا از شناسایی و پاسخ به موقع به تهدیدات احتمالی

اطمینان حاصل گردد.

- **به‌روزرسانی و وصله‌های مدل:**
  - مدل‌های هوش مصنوعی و زیرساخت‌های مربوطه با آخرین وصله‌های امنیتی و بهبودها در مدیریت ورودی‌های خصمانه به‌روز نگه داشته شود.
- **اقدامات اخلاقی و ایمنی:**
  - راهنماهای اخلاقی و ایمنی به مدل هوش مصنوعی ادغام کرده تا از تولید محتوای مضر یا اجرای دستورات غیرمجاز جلوگیری شود.
- **نظارت انسانی:**
  - در برنامه‌های بحرانی، قابلیت‌های نظارت و مداخله انسانی در نظر شود تا خروجی‌های بالقوه مضر از سامانه‌های هوش مصنوعی مورد بررسی و اصلاح قرار گیرد.

۳-۱۴ تزریق غیر مستقیم درخواست

تزریق غیرمستقیم درخواست[1] نوعی از آسیب‌پذیری‌های امنیتی است که زمانی رخ می‌دهد که ورودی‌ها به یک مدل زبان به‌صورت غیرمستقیم از طریق روش‌های دیگر دستکاری می‌شوند، مانند تغییر محیط یا زمینه‌ای که مدل در آن عمل می‌کند. برخلاف تزریق مستقیم درخواست، که شامل ارائه ورودی‌های مخرب به‌طور مستقیم به مدل است، تزریق غیرمستقیم از نحوه ساختاردهی یا ارائه اطلاعات به مدل بدون تغییر مستقیم درخواست ورودی بهره‌برداری می‌کند.

هدف اصلی تزریق غیرمستقیم درخواست فریب دادن یا تأثیرگذاری بر خروجی‌های مدل زبان بدون تعامل مستقیم با رابط ورودی آن است که اغلب باعث می‌شود شناسایی و کاهش این نوع تزریق نسبت به تزریق مستقیم دشوارتر باشد. این نوع تزریق نیاز به درک جامع‌تری از معماری سامانه و فرآیندهای مدیریت داده‌ها دارد تا بتوان به‌طور مؤثری در برابر آن محافظت کرد.

۳-۱۴-۱ ویژگی‌های کلیدی تزریق غیرمستقیم درخواست

- تزریق غیرمستقیم درخواست دارای چندین ویژگی کلیدی است که آن را از تزریق مستقیم درخواست متمایز می‌کند و عمدتاً بر نحوه دستکاری عوامل خارجی برای تأثیر بر خروجی مدل زبان تمرکز دارد:
- **تأثیرات محیطی:**
  - تزریق به تغییر محیط یا زمینه‌ای که مدل زبان در آن عمل می‌کند متکی است، نه تغییر مستقیم درخواست‌هایی که مدل دریافت می‌کند.

1. Indirect Propmt Injection

- **دستکاری منابع داده خارجی:**
  - منابع داده خارجی، پایگاه‌های داده یا مخازن محتوایی که مدل به آن‌ها دسترسی دارد می‌توانند تغییر کنند تا به‌طور غیرمستقیم بر پاسخ‌های مدل تأثیر بگذارند.
- **بهره‌برداری از وابستگی‌های زمینه‌ای:**
  - این نوع تزریق از وابستگی مدل به اطلاعات زمینه‌ای یا پس‌زمینه‌ای که بر درک و خروجی آن تأثیر می‌گذارد سوءاستفاده می‌کند.
- **مسیرهای دسترسی غیرمستقیم:**
  - به جای حمله مستقیم به مدل، تزریق به بهره‌برداری از سایر اجزاء یا سامانه‌هایی می‌پردازد که با مدل تعامل دارند، مانند پردازش داده‌های بالادستی یا نقاط ادغام.
- **پیچیدگی بیشتر:**
  - این نوع تزریق معمولاً شامل راهبردهای پیچیده‌تری است که نیاز به درک عمیق‌تری از معماری سامانه، جریان‌های داده و وابستگی‌ها دارد.
- **سخت‌تر برای شناسایی:**
  - ماهیت غیرمستقیم دستکاری، شناسایی آن را دشوارتر می‌کند، زیرا تغییرات آشکاری در خود درخواست‌های ورودی وجود ندارد.
- **پتانسیل تأثیر گسترده:**
  - دستکاری‌های غیرمستقیم می‌توانند بر چندین حوزه از یک سامانه تأثیر بگذارند و اثرات موجی ایجاد کنند که ممکن است عملکردها یا ماژول‌های مختلفی را که به داده‌های دستکاری‌شده وابسته‌اند، تحت تأثیر قرار دهد.

۳-۱۴-۲ دارایی‌های هدف

- تزریق غیرمستقیم درخواست به دارایی‌های خاصی هدف‌گیری می‌کند که برای عملیات و امنیت یک مدل زبان و سامانه‌های اطراف آن حیاتی هستند. برخی از دارایی‌های اصلی هدف شامل موارد زیر است:
- **منابع داده خارجی:**
  - پایگاه‌های داده، APIها یا مخازن محتوایی که مدل زبان برای استخراج اطلاعات به آن‌ها دسترسی دارد. با دستکاری این منابع داده، یک مهاجم می‌تواند به‌طور غیرمستقیم بر خروجی‌های مدل تأثیر بگذارد.

- **زمینه و پیکربندی سامانه:**
  - تنظیمات، پیکربندی‌ها و متغیرهای محیطی که ممکن است نحوه پردازش ورودی‌ها و تولید خروجی‌ها را تغییر دهند. این شامل متاداده یا اطلاعات زمینه‌ای است که بر رفتار مدل تأثیر می‌گذارد.
- **اجزای ارتباطی:**
  - سایر اجزای سامانه یا برنامه‌هایی که با مدل زبان تعامل دارند یا ورودی‌هایی به آن ارائه می‌دهند. این می‌تواند شامل پردازش‌های پیشینه یا میانی باشد که داده‌ها را برای مدل آماده می‌کند.
- **پروفایل‌های کاربری و داده‌های تاریخی:**
  - داده‌های مربوط به رفتار کاربران، ترجیحات یا تعاملات قبلی که مدل از آن‌ها برای سفارشی‌سازی پاسخ‌ها استفاده می‌کند. دستکاری این داده‌ها می‌تواند به‌طور غیرمستقیم نتایج را کج کند.
- **سازوکارهای امنیتی و کنترل دسترسی:**
  - سازوکارهایی که جریان داده و مجوزهای دسترسی را درون سامانه مدیریت می‌کنند. بهره‌برداری از آسیب‌پذیری‌ها در این زمینه می‌تواند راه‌هایی برای تزریق غیرمستقیم فراهم کند.
- **گزارش‌های عملیاتی و ابزارهای نظارتی:**
  - گزارش‌ها یا سامانه‌های نظارتی که عملکرد و ورودی‌های مدل را پیگیری می‌کنند. دستکاری غیرمستقیم می‌تواند شامل تغییر در نحوه درک و گزارش رفتار مدل توسط این سامانه‌ها باشد.
- **داده‌های آموزش و بهینه‌سازی:**
  - مجموعه‌های داده‌ای که برای آموزش یا بهینه‌سازی مدل استفاده می‌شوند، می‌توانند هدف قرار گیرند تا سوگیری‌ها یا نادرستی‌هایی را که در پاسخ‌های مدل باقی می‌ماند، معرفی کنند و بر رفتار آینده آن تأثیر بگذارند.

۳-۱۴-۳ آسیب‌پذیری‌ها

تزریق غیرمستقیم درخواست از آسیب‌پذیری‌های خاصی درون اکوسامانه گسترده‌تر سامانه و داده‌ها که از یک مدل زبان پشتیبانی می‌کند، استفاده می‌کند. درک این آسیب‌پذیری‌ها برای توسعه دفاع‌های مؤثر بسیار مهم است. در اینجا آسیب‌پذیری‌های کلیدی مرتبط با تزریق غیرمستقیم درخواست آورده شده است:

- **یکپارچگی داده‌های خارجی:**
  - عدم اعتبارسنجی و تأیید دقیق منابع داده خارجی که مدل به آن‌ها وابسته است، می‌تواند منجر به دستکاری شود و بر خروجی‌های مدل تأثیر بگذارد.
- **وابستگی‌های زمینه‌ای:**
  - وابستگی مدل به اطلاعات زمینه‌ای یا پس‌زمینه‌ای (مانند متاداده یا پیکربندی‌ها) می‌تواند یک آسیب‌پذیری باشد اگر آن زمینه‌ها قابل تأثیرگذاری یا خراب شدن باشند.
- **کنترل‌های دسترسی ضعیف:**
  - محدودیت‌های ناکافی بر روی منابع داده، فایل‌های پیکربندی، یا سامانه‌های رابطدهنده می‌تواند به مهاجمان اجازه دهد تا به‌طور غیرمجاز دستکاری کنند.
- **ضعف در اجزای رابطدهنده:**
  - آسیب‌پذیری‌ها در برنامه‌ها یا میانی که داده‌ها را به مدل ارائه می‌دهند می‌تواند به‌عنوان مسیرهای غیرمستقیم برای تزریق عمل کند اگر بتوانند مورد بهره‌برداری قرار بگیرند.
- **نظارت ناکافی:**
  - عدم وجود سامانه‌های نظارتی قوی که بتوانند فعالیت‌های غیرعادی و تغییرات در جریان داده‌ها یا پیکربندی‌ها را شناسایی کنند، اجازه می‌دهد تا ناهنجاری‌ها نادیده گرفته شوند.
- **ضعف در مدیریت منشا داده:**
  - پیگیری ضعیف از منشا داده‌ها و تغییرات آن‌ها در طول زمان، شناسایی و تأیید یکپارچگی داده‌های استفاده‌شده توسط مدل را دشوار می‌سازد.
- **مسائل مدیریت پیکربندی:**
  - نقاط ضعف در نحوه مدیریت و تأمین امنیت پیکربندی‌های محیطی می‌تواند راه‌هایی را برای مهاجمان فراهم کند تا تنظیماتی را که بر عملیات مدل تأثیر می‌گذارد، تغییر دهند.
- **آسیب‌پذیری‌های داده‌های آموزشی:**
  - قرار گرفتن در معرض مجموعه‌های داده آموزشی و بهینه‌سازی شده که به‌طرز فاسد یا سوگیری شده‌اند، می‌تواند به‌طور غیرعمدی سوگیری‌ها یا آسیب‌پذیری‌های پایداری را در رفتار مدل معرفی کند.

۳-۱۴-۴ سناریو تهدید

یک شرکت یک چت‌بات را به سامانه‌های داخلی خود اضافه می‌کند که برای کمک به کاربران در

پاسخ به سؤالات و ارائه راهنمایی بر اساس پروفایل‌های مشتری و داده‌های تعامل تاریخی طراحی شده است. یک مهاجم می‌خواهد با هدف قرار دادن داده‌ها و سامانه‌هایی که بر خروجی‌های آن تأثیر می‌گذارند که به‌طور غیرمستقیم پاسخ‌های چت‌بات را دستکاری کند تا به شهرت شرکت آسیب بزند یا اطلاعات حساس کاربران را استخراج کند،

- **جمع‌آوری اطلاعات:**
  - مهاجم اطلاعات دقیقی درباره معماری سامانه جمع‌آوری می‌کند و بر درک نحوه ادغام چت‌بات با پایگاه‌های داده خارجی تمرکز می‌کند. این شامل شناسایی مسیرهایی است که چت‌بات از طریق آن‌ها داده‌ها را استخراج می‌کند.
- **بهره‌برداری از سامانه‌های خارجی:**
  - مهاجم راهی برای دسترسی به پایگاه‌های داده خارجی یا سامانه‌های مدیریت پروفایل پیدا می‌کند، چه از طریق سازوکارهای کنترل دسترسی ضعیف و چه از طریق فیشینگ برای کسب اعتبار از کارکنان بی‌خبر.
- **دستکاری داده:**
  - پس از به دست آوردن دسترسی، مهاجم شروع به تغییر نقاط کلیدی داده در تاریخچه‌های تعامل می‌کند و اطلاعات گمراه‌کننده یا نامناسب را تزریق می‌کند که در طول تعاملات توسط چت‌بات استخراج می‌شود.
- **تغییر زمینه:**
  - مهاجم اطلاعات زمینه‌ای یا تنظیمات پیکربندی سامانه، مانند تنظیمات پیش‌فرض پاسخ‌ها یا اولویت منابع داده را تغییر می‌دهد که بر منطق چت‌بات در تولید پاسخ‌ها تأثیر می‌گذارد.
- **نظارت مستمر:**
  - مهاجم تعاملات با چت‌بات را نظارت می‌کند و تغییرات خود را اصلاح می‌کند تا به‌طور مؤثر پاسخ‌های چت‌بات را به گونه‌ای که با اهدافش همسو باشد، تغییر دهد و موجب اطلاعات نادرست یا نقض حریم خصوصی گردد.

۵-۱۴-۳ نشانه‌های احتمال نفوذ

نشانه‌های احتمال نفوذ برای تزریق غیرمستقیم درخواست در چت‌بات‌ها می‌توانند به شناسایی زمانی که یک سامانه به‌طرز بالقوه‌ای دستکاری شده است، کمک کنند. در اینجا برخی از کلیدی که باید به آن‌ها توجه کرد، آورده شده است:

- **تغییرات غیرمنتظره در خروجی:**
  - چت‌بات پاسخ‌هایی تولید می‌کند که غیرمعمول، نادرست یا بی‌ربط به زمینه هستند و با رفتار معمول آن تفاوت دارند.
- **تجربه کاربری ناپایدار:**
  - کاربران گزارش‌هایی از ناهماهنگی‌ها یا عدم تناقض در تعاملات ارائه می‌دهند که نشان می‌دهد چت‌بات برای یک پرسش خاص، پاسخ‌های متفاوتی ارائه می‌دهد.
- **الگوهای غیرعادی دسترسی به داده:**
  - الگوهای غیرمعمول دسترسی به پایگاه‌های داده یا منابع خارجی که چت‌بات به آن‌ها وابسته است، مانند افزایش فراوانی یا دسترسی خارج از پارامترهای معمول.
- **تغییرات در پیکربندی سامانه:**
  - تغییرات غیرمنتظره در تنظیمات سامانه یا متاداده که بر نحوه اولویت‌بندی یا پردازش اطلاعات توسط چت‌بات تأثیر می‌گذارد.
- **افزایش نرخ خطا:**
  - افزایش قابل توجهی در خطاها یا تعاملات ناموفق در سامانه چت‌بات، که ممکن است نشان‌دهنده دستکاری داده‌ها یا منطق آن باشد.
- **ترافیک شبکه غیرعادی:**
  - فعالیت‌های نامعمول شبکه، مانند دسترسی به داده‌ها از آدرس‌های IP ناشناخته یا مشکوک، می‌تواند نشان‌دهنده تلاش‌های غیرمجاز برای دستکاری منابع داده باشد.
- **ناهماهنگی‌ها در لاگ‌های داده:**
  - فایل‌های لاگ نشان‌دهنده ناهنجاری‌هایی مانند تغییرات در داده‌ها یا لاگ‌های دسترسی که با الگوهای استفاده مورد انتظار همخوانی ندارند.
- **گزارش‌های افشای اطلاعات حساس:**
  - مواردی که در آن‌ها اطلاعات حساس کاربران در طول تعاملات چت‌بات گزارش می‌شود، نشان‌دهنده احتمال دستکاری در مجوزهای دسترسی یا مسیرهای داده است.
- **شکایات یا گزارش‌های کاربران:**
  - افزایش شکایات یا گزارش‌های کاربران که رفتارهای عجیب یا اطلاعات نادرست را از چت‌بات توصیف می‌کنند.

۶-۱۴-۳ تأثیرات

- **نقص یکپارچگی داده:**
  - دستکاری منابع داده خارجی یا ورودی‌های زمینه‌ای می‌تواند منجر به تولید خروجی‌های نادرست یا گمراه‌کننده شود و در نتیجه یکپارچگی اطلاعات ارائه‌شده توسط سامانه هوش مصنوعی را زیر سؤال ببرد.
- **کاهش اعتماد:**
  - اگر کاربران پاسخ‌های نادرست یا نامناسب دریافت کنند، اعتماد به سامانه هوش مصنوعی و سازمان پشت آن به شدت تضعیف می‌شود که این موضوع ممکن است بر مشارکت و وفاداری کاربران تأثیر بگذارد.
- **آسیب به شهرت:**
  - انتشار اطلاعات نادرست یا داده‌های حساس می‌تواند به شهرت یک سازمان آسیب بزند و منجر به چالش‌های روابط عمومی شود و ممکن است مشتریان را به سمت رقبای دیگر سوق دهد.
- **اختلالات عملیاتی:**
  - خروجی‌های نادرست ممکن است منجر به ناکارآمدی‌های عملیاتی شوند، همچون اتخاذ تصمیمات اشتباه، تخصیص نادرست منابع یا شکست در فرآیندهای خودکار که به بینش‌های تولیدشده توسط هوش مصنوعی وابسته هستند.
- **زیان‌های مالی:**
  - خروجی‌های گمراه‌کننده ممکن است عواقب مالی به دنبال داشته باشد، از هزینه‌های فوری مربوط به اصلاح خطاها تا تأثیرات بلندمدت مانند از دست دادن کسب‌وکار یا کاهش قیمت سهام به دلیل آسیب به شهرت.
- **ریسک‌های قانونی و انطباق:**
  - انتشار ناخواسته اطلاعات حساس یا خصوصی ممکن است منجر به نقض مقررات حفاظت از داده‌ها گردد و به اقدامات قانونی، جریمه‌ها و نیاز به تلاش‌های پرهزینه برای جبران آسیب منجر شود.
- **آسیب‌پذیری‌های امنیتی:**
  - اگر یک مهاجم بتواند خروجی‌ها را از طریق تزریق غیرمستقیم درخواست دستکاری کند، این می‌تواند آسیب‌پذیری‌های بیشتری در سامانه را نمایان کند که ممکن است برای حملات

اضافی مورد بهره‌برداری قرار گیرد.

- **پخش اطلاعات نادرست:**
  - خروجی‌های تغییر یافته می‌توانند به گسترش اطلاعات نادرست کمک کنند، که می‌تواند تأثیرات اجتماعی وسیع‌تری داشته باشد، به‌ویژه در حوزه‌های حیاتی مانند اخبار، سلامت و مالی.

۳-۱۴-۷ راه‌های مقابله

کاهش ریسک‌های مرتبط با تزریق غیرمستقیم درخواست مستلزم پیاده‌سازی مجموعه‌ای جامع از راهبردها است که بر حفظ یکپارچگی داده، معماری سامانه و فرآیندهای عملیاتی تمرکز دارند. در اینجا برخی از تدابیر مؤثر کاهش خطر آورده شده است:

- **اعتبارسنجی داده و بررسی یکپارچگی:**
  - به‌طور منظم یکپارچگی منابع داده خارجی تأیید و اعتبارسنجی شود. از سازوکارهای بررسی، مانند امضای دیجیتال یا سایر روش‌های تأیید استفاده شده تا اصالت و یکپارچگی داده‌ها قبل از استفاده توسط مدل زبان تضمین شود.
- **کنترل‌های دسترسی قوی:**
  - کنترل‌های سختگیرانه دسترسی و سازوکارهای احراز هویت برای تمامی پایگاه‌داده‌ها و سامانه‌هایی که با هوش مصنوعی تعامل دارند اعمال گردد. اطمینان حاصل شود که تنها کارکنان معتبر به داده‌های حساس و تنظیمات پیکربندی دسترسی دارند.
- **بازرسی‌های امنیتی منظم:**
  - بازرسی‌های امنیتی و آزمایش‌های نفوذ را به‌طور مکرر انجام داده تا نقاط ضعف موجود در معماری سامانه را شناسایی و برطرف شود، به ویژه بر روی جریان‌های داده و نقاط ادغام تمرکز گردد.
- **سامانه‌های نظارت جامع:**
  - ابزارهای نظارتی و لاگ‌گذاری پیاده‌سازی شود تا دسترسی به منابع داده و تغییرات در پیکربندی‌های سامانه پیگیری گردد. از تکنیک‌های تشخیص ناهنجاری برای علامت‌گذاری فعالیت‌های غیرعادی یا دستکاری‌های داده استفاده شود.
- **منشا داده و قابلیت ردیابی:**
  - سوابق دقیقی از منبع داده و تغییرات آن در طول زمان حفظ گردد تا به شناسایی هرگونه دستکاری بالقوه و پاسخ سریع به آن‌ها کمک کند.

- **معماری داده‌های تقسیم‌شده:**
  - از تقسیم و تفکیک داده‌ها استفاده گردد تا تأثیر هر منبع داده خاصی که ممکن است به خطر بیفتد، محدود شود و از تأثیرگذاری مهاجمان بر کل سامانه جلوگیری به عمل آید.
- **بررسی یکپارچگی زمینه‌ای:**
  - سامانه‌ها و سیاست‌هایی پیاده‌سازی شود که به‌طور مداوم صحت و مناسب بودن اطلاعات زمینه‌ای که توسط مدل زبان استفاده می‌شود را تأیید کنند تا خروجی‌ها دقیق و مرتبط باشد.
- **برنامه‌ریزی پاسخ به حوادث:**
  - یک برنامه پاسخ به حوادث توسعه داده و به‌طور منظم به‌روزرسانی شود که شامل رویه‌های خاص برای مقابله با حملات تزریق غیرمستقیم درخواست باشد و امکان حل سریع مسائل را فراهم کند.
- **آموزش و آگاهی کاربران:**
  - کارکنان و کاربران را برای شناسایی نشانه‌های بالقوه دستکاری آموزش داده و آن‌ها به گزارش هرگونه فعالیت مشکوک ترغیب شوند تا فرهنگ آگاهی امنیتی در سازمان تقویت گردد.
- **به‌روزرسانی و وصله‌های منظم سامانه:**
  - اطمینان حاصل شود که همه اجزای سامانه هوش مصنوعی و زیرساخت‌های مربوطه با آخرین وصله‌ها و به‌روزرسانی‌های امنیتی نگه‌داری شوند تا در برابر آسیب‌پذیری‌های شناخته شده محافظت شوند.

۴ مطالعه موردی شناسایی و مقابله با حملات مبتنی بر هوش‌مصنوعی در حوزه‌های کاربردی منتخب

۴-۱ مقدمه

در این فصل به منظور کاربردی سازی مطالعات انجام شده، حملات شناسایی و مشخص‌شده در فصول ۲ و ۳ در ۵ حوزه کاربردی مالی و بانکی، سلامت، پیام‌رسان‌های همراه، وبگاه‌های خبری، مدل‌های زبانی بزرگ به صورت موردی، بررسی شده و حملات تهاجمی و تخاصمی که در حوزه مربوطه معنادار بوده به همراه دارایی‌های هدف، تأثیر و راهکار مقابله با آن‌ها مشخص شده‌اند.

۴-۲ شناسایی و مقابله با حملات مبتنی بر هوش مصنوعی در حوزه مالی بانکی

جدول ۵: حملات هوش مصنوعی تهاجمی در حوزه مدل‌های زبانی بزرگ

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تحلیل ترافیک شبکه	دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های مدیریت منابع انسانی، سامانه‌های گزارش‌دهی و ریسک دارایی نامشهود موارد: مالکیت فکری، روابط با مشتریان و شرکا.	افشای الگوی ارتباطات و رفتار سیستم	رمزنگاری end-to-end، تحلیل متادیتا با هوش مصنوعی ،تطبیق رفتار شبکه
تشخیص آسیب‌پذیری	دارایی‌های مالی موارد: ذخایر سرمایه‌ای دارایی‌های زیرساختی و اجزاء شبکه موارد: سخت‌افزارهای بانکی، نرم‌افزار بانکداری متمرکز، شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات، تجهیزات پردازش تراکنش و سرورهای مالی، تجهیزات ذخیره‌سازی، سامانه‌های فناوری اطلاعات، سامانه‌های نظارتی و انطباق سامانه‌های امنیت سایبری، دستگاه‌های خودپرداز، مراکز داده دارایی‌های زیرساخت فناوری اطلاعات موارد: سرورها، دستگاه‌های شبکه، پشتیبان‌گیری و بازیابی، زیرساخت ابری،پلتفرم‌های مدیریت مشتری، پلتفرم‌های سرمایه‌گذاری، سامانه‌های پردازش تراکنش، سامانه‌های مدیریت منابع انسانی، سامانه‌های حقوقی و انطباق، سامانه‌های گزارش‌دهی و ریسک، پلتفرم‌های آموزش و توسعه، پلتفرم‌های وام و اعتبار، پلتفرم‌های تحلیل داده و هوش تجاری	شناسایی نقاط ضعف برای سوءاستفاده برای حمله بعدی	آزمون‌های منظم آسیب‌پذیری، محدود کردن دسترسی به سیستم‌های حساس اسکن امنیتی منظم تحلیل کد ایستا/پویا Patch Management Bug Bounty

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تشخیص آسیب‌پذیری	دارایی نامشهود موارد: برند و اعتبار بانک، مالکیت فکری، روابط با مشتریان و شرکا.، مجوزهای قانونی و نظارتی دارایی داده و حاکمیت داده موارد: داده‌های مشتری، رضایت و حریم خصوصی مشتری، حاکمیت و مدیریت داده،داده‌های شرکای تجاری، داده‌های طرف سوم،داده‌های کارکنان، شاخص‌های مبتنی بر اینترنت، مجموعه داده‌های تجاری، داده‌های نظارتی و انطباقی اطلاعات مشتری، داده‌های بازار، شاخص‌های بازار مالی دارایی خدمات موارد: خدمات بانکداری دیجیتال،خدمات پرداخت، خدمات سرمایه‌گذاری، خدمات بیمه‌ای مرتبط با بانکداری دارایی‌های امنیت فناوری اطلاعات موارد: امنیت شبکه، سیستم‌های رمزنگاری، راهکارهای مدیریت هویت و دسترسی (IAM)، حفاظت از نقاط پایانی، سیستم‌های SIEM، مدیریت آسیب‌پذیری‌ها، پاسخ به حوادث امنیتی راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی، بانکداری همراه، کیف پول دیجیتال، راهکارهای نئوبانک، مدیریت API سیستم‌های حفاظت از داده و انطباق موارد: حفاظت از داده‌ها، مدیریت انطباق، کنترل دسترسی و حریم خصوصی، مدیریت حوادث و فارنزیک، پشتیبان‌گیری و بازیابی دارایی‌های عملیاتی موارد: پلتفرم GRC، حاکمیت،مدیریت ریسک، مدیریت حوادث، کنترل‌های داخلی	شناسایی نقاط ضعف برای سوءاستفاده برای حمله بعدی	آزمون‌های منظم آسیب‌پذیری، محدود کردن دسترسی به سیستم‌های حساس اسکن امنیتی منظم تحلیل کد ایستا/پویا Patch Management Bug Bounty
جاسوسی	دارایی‌های زیرساختی و اجزاء شبکه موارد: شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات، سامانه‌های امنیت سایبری دارایی خدمات موارد: خدمات وام و اعتبار دارایی‌های امنیت فناوری اطلاعات موارد: امنیت شبکه، سیستم‌های SIEM، پاسخ به حوادث امنیتی	نظارت غیرمجاز بر داده‌های حساس دسترسی غیرمجاز به داده محرمانه	رمزنگاری ارتباطات، پنهان کردن اطلاعات موقعیت‌یابی (مانند IP مجازی) جلوگیری از خروج اطلاعات (DLP) کنترل رسانه‌های ذخیره‌سازی رفتارشناسی کاربر (UEBA)

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
پروفایل‌سازی هوشمند هدف‌محور	دارایی نامشهود موارد: روابط با مشتریان و شرکا. دارایی داده و حاکمیت داده موارد: شاخص‌های مبتنی بر اینترنت، داده‌های تحلیلی و پیش‌بینی	سوءاستفاده از پروفایل‌های کاربری برای حملات هدفمند سوءاستفاده هدفمند از اطلاعات فردی	محدود کردن جمع‌آوری داده‌های شخصی، کنترل دسترسی به داده‌های مشتریان ناشناس‌سازی داده، محدودیت دسترسی، تحلیل رفتاری نظارت مستمر بر دسترسی به اطلاعات حساس آموزش کارکنان برای شناسایی و گزارش فعالیت‌های مشکوک پیاده‌سازی سامانه‌های تشخیص رفتار غیرعادی کاربران (User Behavior Analytics)
تحلیل رفتاری برای یافتن راه‌های جدید سوءاستفاده	دارایی‌های داده و حاکمیت داده موارد: شاخص‌های مبتنی بر اینترنت	کشف آسیب‌پذیری‌های جدید از طریق الگوهای رفتاری شناسایی نقاط ضعف از رفتار کاربران	آنالیز ترافیک شبکه برای شناسایی حملات، محدود کردن دسترسی به داده‌های حساس UEBA سیاست پویا Obfuscation ساختاری آموزش امنیتی SIEM و لاگ‌گیری پیشرفته
پیش‌بینی نتایج حملات	دارایی‌های زیرساختی و اجزاء شبکه موارد: تجهیزات ذخیره‌سازی، سامانه‌های فناوری اطلاعات دارایی‌های زیرساخت فناوری اطلاعات موارد: راهکارهای ذخیره‌سازی، زیرساخت ابری دارایی‌های امنیت فناوری اطلاعات موارد: امنیت شبکه،حفاظت از نقاط پایانی	افزایش اثربخشی حملات آینده بهینه‌سازی حمله براساس پاسخ سیستم	آنالیز رفتاری داده‌ها برای شناسایی الگوهای غیرطبیعی، آمادگی برای مقابله با حملات آینده شبیه‌سازی MITRE ATT&CK Threat Modeling Reverse ML Predictive Analytics Risk Assessment
رمزشکنی	دارایی خدمات موارد: خدمات بیمه‌ای مرتبط با بانکداری دارایی‌های امنیت فناوری اطلاعات موارد: سیستم‌های رمزنگاری سیستم‌های حفاظت از داده و انطباق دارایی داده و حاکمیت داده موارد: کنترل دسترسی و حریم خصوصی	دسترسی غیرمجاز به داده‌های رمزگذاری‌شده افشای داده‌های رمزگذاری‌شده	استفاده از روش‌های قوی احراز هویت مانند :محدود کردن تلاش‌های ورود Perfect Forward Secrecy مقابله با Side-Channel Attacks رمزنگاری مقاوم کوانتومی مدیریت کلید امن (HSM)

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
کش‌کاوی	دارایی داده و حاکمیت داده موارد: شاخص‌های بازار مالی، داده‌های بازار، داده‌های تحلیلی و پیش‌بینی، شاخص‌های مبتنی بر اینترنت دارایی‌های امنیت فناوری اطلاعات موارد: حفاظت از نقاط پایانی سیستم‌های حفاظت از داده و انطباق موارد: حفاظت از داده‌ها دارایی‌های عملیاتی موارد: مدیریت ریسک	استخراج غیرمجاز داده‌های کش‌شده (cache)	رمزنگاری داده‌های ذخیره‌شده، مخفی کردن اطلاعات شخصی در سیستم‌های کش کنترل دسترسی به داده بررسی خروجی مدل الگوریتم‌های مقاوم حریم خصوصی
مرد میانی	دارایی‌های زیرساختی و اجزاء شبکه موارد: شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات، سامانه‌های امنیت سایبری دارایی‌های زیرساخت فناوری اطلاعات موارد: دستگاه‌های شبکه، پلتفرم‌های تحلیل داده و هوش تجاری دارایی‌های امنیت فناوری اطلاعات (A۰۷) موارد: امنیت شبکه	دستکاری/شنود داده‌های در حال انتقال	استفاده از TLS/SSL، امضای دیجیتال برای تأیید منشأ داده‌ها رمزنگاری TLS اعتبارسنجی دوطرفه HSTS استفاده از VPN
تولید خودکار دامنه	دارایی‌های زیرساختی و اجزاء شبکه موارد: سخت‌افزارهای بانکی، دستگاه‌های خودپرداز	پنهان‌کردن سرورهای فرمان‌کنترل (C2) ایجاد زیرساخت حملات گسترده (botnet)	فیلتر کردن دامنه‌های مشکوک، اعتبارسنجی دامنه‌ها با DNSSEC بلاک لیست DNS، هوش تهدید، فیلتر دامنه بررسی الگوی نام دامنه فیلتر DGA تهدیدشناسی دامنه‌ها
مبهم‌سازی بدافزار	دارایی‌های زیرساختی و اجزاء شبکه موارد: سخت‌افزارهای بانکی، نرم‌افزار بانکداری متمرکز، شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات، تجهیزات پردازش تراکنش و سرورهای مالی، سامانه‌های فناوری اطلاعات، سامانه‌های امنیت سایبری، دستگاه‌های خودپرداز، مراکز داده دارایی‌های زیرساخت فناوری اطلاعات موارد: سرورها، زیرساخت ابری، پلتفرم‌های سرمایه‌گذاری، سامانه‌های پردازش تراکنش دارایی داده و حاکمیت داده موارد: داده‌های شرکای تجاری، داده‌های طرف سوم، شاخص‌های مبتنی بر اینترنت، مجموعه داده‌های تجاری دارایی خدمات موارد: خدمات بانکداری دیجیتال، خدمات پرداخت دارایی‌های امنیت فناوری اطلاعات موارد: امنیت شبکه راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی، بانکداری همراه، کیف پول دیجیتال، راهکارهای نئوبانک، مدیریت API	فرار از سیستم‌های تشخیص با پنهان‌کردن کد پنهان‌سازی بدافزار از آنتی‌ویروس	استفاده از سیستم‌های تشخیص نفوذ، فیلتر کردن ترافیک شبکه تحلیل استاتیک و دینامیک شبیه‌سازی حملات De-obfuscation Techniques

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تولید داده‌های مصنوعی	دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های آموزش و توسعه دارایی داده و حاکمیت داده موارد: شاخص‌های مبتنی بر اینترنت، داده‌های تحلیلی و پیش‌بینی	مسموم‌سازی مدل‌های یادگیری ماشین گمراه‌سازی مدل‌های تحلیلی یا یادگیری ماشین	اعتبارسنجی داده‌ها با الگوریتم‌های هوش مصنوعی، محدود کردن دسترسی به داده‌های واقعی اعتبارسنجی ورودی، تحلیل آماری، کشف ناهنجاری آموزش مدل‌های تشخیص ناهنجاری برای شناسایی داده‌های غیرواقعی اعتبارسنجی دقیق منابع داده‌ای و بررسی صحت و یکپارچگی داده‌ها (Data Integrity Checks) استفاده از تکنیک‌های شناسایی داده‌های جعلی (Data Provenance) نظارت مستمر بر کیفیت داده‌ها (Data Quality Monitoring)
تولید هوشمند نظرات جعلی	دارایی‌های زیرساختی و اجزاء شبکه موارد: شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات، سامانه‌های فناوری اطلاعات، سامانه‌های امنیت سایبری، دستگاه‌های خودپرداز، مراکز داده دارایی‌های زیرساخت فناوری اطلاعات موارد: دستگاه‌های شبکه، دارایی داده و حاکمیت داده موارد: مجموعه داده‌های تجاری	دستکاری اعتماد کاربران فریب کاربران یا الگوریتم‌های تصمیم‌گیری	محدود کردن دسترسی به داده‌های کاربران، اعتبارسنجی نظرات با هوش مصنوعی تحلیل احساسات با هوش مصنوعی تشخیص spam CAPTCHA رفتاری
حمله به CAPTCHA	دارایی‌های زیرساختی و اجزاء شبکه موارد: نرم‌افزار بانکداری متمرکز دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های آموزش و توسعه دارایی داده و حاکمیت داده موارد: شاخص‌های مبتنی بر اینترنت دارایی خدمات موارد: خدمات بانکداری دیجیتال دارایی‌های امنیت فناوری اطلاعات موارد: راهکارهای مدیریت هویت و دسترسی (IAM) راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی، بانکداری همراه، کیف پول دیجیتال، راهکارهای نئوبانک	سوءاستفاده خودکار از خدمات عبور خودکار از سیستم‌های ضد ربات	استفاده از CAPTCHAهای هوشمند، فیلتر کردن ترافیک رباتیک CAPTCHA رفتاری و پویا بررسیfingerprint مرورگر/دستگاه شواهد تعامل انسان (Human Inter-action Proof)
تولید خودکار اطلاعات نادرست	دارایی‌های زیرساختی و اجزاء شبکه موارد: سخت‌افزارهای بانکی، دستگاه‌های خودپرداز، دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های مدیریت منابع انسانی داده و حاکمیت داده موارد: حاکمیت و مدیریت داده	گمراه‌سازی تحلیل‌گر یا کاربران	تحلیل محتوا با هوش مصنوعی تشخیص زبان طبیعی غیرعادی بررسی الگوی نگارشی

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
سیبل	دارایی‌های زیرساختی و اجزاء شبکه موارد: نرم‌افزار بانکداری متمرکز- تجهیزات پردازش تراکنش و سرورهای مالی دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های مدیریت مشتری، پلتفرم‌های سرمایه‌گذاری، سامانه‌های پردازش تراکنش، سامانه‌های مدیریت منابع انسانی، پلتفرم‌های وام و اعتبار دارایی خدمات موارد: خدمات بانکداری دیجیتال، خدمات پرداخت، خدمات وام و اعتبار دارایی‌های امنیت فناوری اطلاعات موارد: سیستم‌های رمزنگاری، راهکارهای مدیریت هویت و دسترسی (IAM)، حفاظت از نقاط پایانی، سیستم‌های SIEM، مدیریت آسیب‌پذیری‌ها، پاسخ به حوادث امنیتی راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی، بانکداری همراه،کیف پول دیجیتال، راهکارهای نئوبانک	اختلال در شبکه از طریق هویت‌های جعلی	احراز هویت منحصر به فرد برای هر گره، ثبت فعالیت‌های گره‌ها تحلیل رفتار بیومتریک، Liveness Detection، بررسی داده‌های غیرعادی، زیست‌سنجی،
فیشینگ هدفمند	دارایی‌های زیرساختی و اجزاء شبکه موارد: شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات، سامانه‌های امنیت سایبری دارایی‌های زیرساخت فناوری اطلاعات موارد: دستگاه‌های شبکه، پلتفرم‌های تحلیل داده و هوش تجاری دارایی داده و حاکمیت داده موارد: شاخص‌های مبتنی بر اینترنت، امنیت شبکه، داده (اعتبارنامه‌ها) دارایی‌های امنیت فناوری اطلاعات موارد: امنیت شبکه دارایی خدمات موارد:خدمات ایمیل، خدمات پیامک، خدمات بانکداری دیجیتال، خدمات پرداخت	سرقت اطلاعات حساس (مثل رمزها) فریب کاربر برای افشای اطلاعات	فیلتر کردن ایمیل‌های مشکوک، آموزش کاربران برای تشخیص ایمیل‌های جعلی بررسی ایمیل با هوش مصنوعی آموزش کاربر فیلتر دامنه جعلی DMARC/DKIM/SPF
حدس رمز عبور	دارایی‌های مالی موارد: سرمایه نقدی و سپرده‌ها دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های مدیریت منابع انسانی، دارایی خدمات موارد: خدمات بانکداری دیجیتال دارایی‌های امنیت فناوری اطلاعات موارد: راهکارهای مدیریت هویت و دسترسی (IAM) راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی	سرقت اعتبارنامه‌های کاربری دسترسی غیرمجاز به سیستم	ترکیب رمز با احراز هویت دو مرحله‌ای مخفی کردن الگوهای رمز پیاده‌سازی احراز هویت چندعاملی الزام به سیاست‌های قوی رمز عبور (پیچیدگی، دوره تعویض) قفل‌کردن حساب‌ها پس از تعداد مشخصی تلاش ناموفق نظارت مستمر و هشدار خودکار برای تلاش‌های ناموفق ورود

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
سرقت رمز عبور	دارایی‌های مالی موارد: ذخایر سرمایه‌ای دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های حقوقی و انطباق، پلتفرم‌های وام و اعتبار دارایی نامشهود موارد: مالکیت فکری دارایی داده و حاکمیت داده موارد: شاخص‌های مبتنی بر اینترنت، داده‌های بازار، شاخص‌های بازار مالی دارایی خدمات موارد: خدمات وام و اعتبار، خدمات سرمایه‌گذاری سیستم‌های حفاظت از داده و انطباق دارایی‌های امنیت فناوری اطلاعات موارد: حفاظت از داده‌ها، مدیریت انطباق، کنترل دسترسی و حریم خصوصی دارایی‌های عملیاتی موارد: پلتفرم GRC، حاکمیت، مدیریت ریسک	دسترسی غیرمجاز به حساب‌های کاربری	رمزنگاری رمزها در حین انتقال، محدود کردن دسترسی به فایل‌های رمز پیاده‌سازی احراز هویت چندعاملی بررسی نشت رمز از طریق threat intelligence تشخیص رفتار مشکوک ورود تغییر اجباری رمز دوره‌ای
بروت فورس رمز عبور	دارایی‌های زیرساختی و اجزاء شبکه موارد: سخت‌افزارهای بانکی، نرم‌افزار بانکداری متمرکز، دستگاه‌های خودپرداز، مراکز داده دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های آموزش و توسعه، سامانه‌های مدیریت منابع انسانی دارایی داده و حاکمیت داده موارد: داده‌های کارکنان، اطلاعات مشتری دارایی خدمات موارد: خدمات بانکداری دیجیتال دارایی‌های امنیت فناوری اطلاعات موارد: راهکارهای مدیریت هویت و دسترسی (IAM) راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی، راهکارهای نئوبانک	سرقت اعتبارنامه‌های کاربری دسترسی غیرمجاز از طریق آزمون رمز	استفاده از رمزهای پیچیده، قفل کردن حساب بعد از تعداد تلاش‌های ناموفق محدودیت تلاش ورود CAPTCHA احراز هویت چندعاملی الگوریتم‌های هش قوی مثل bcrypt
جعل	دارایی‌های زیرساختی و اجزاء شبکه موارد: نرم‌افزار بانکداری متمرکز، تجهیزات پردازش تراکنش و سرورهای مالی دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های سرمایه‌گذاری، سامانه‌های پردازش تراکنش، پلتفرم‌های تحلیل داده و هوش تجاری دارایی خدمات موارد: خدمات پرداخت راه‌حل های نوین بانکداری موارد: راهکارهای نئوبانک، مدیریت API	جعل هویت دستگاه/کاربر ساخت داده یا پیام جعلی	احراز هویت بیومتریک، ثبت امضای دیجیتال برای اثبات منشأ Liveness Detection به‌روزرسانی الگوریتم تشخیص ذخیره امن داده زیست‌سنجی

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
جعل بیومتریک	دارایی‌های امنیت فناوری اطلاعات موارد: راهکارهای مدیریت هویت و دسترسی (IAM) دارایی‌های زیرساخت فناوری اطلاعات موارد: سیستم‌های احراز هویت بیومتریک دارایی داده و حاکمیت داده موارد: داده‌های مشتری، داده (الگوهای بیومتریک)،	دسترسی غیرمجاز از طریق بیومتریک جعلی دور زدن سیستم‌های احراز هویت	استفاده از روش‌های چندعاملی، اعتبارسنجی الگوهای بیومتریک استفاده از تأیید هویت چندعاملی Multi-factor Authentication -) (MFA به‌روزرسانی مستمر و ارزیابی امنیتی سامانه‌های بیومتریک
سرقت هویت	دارایی‌های زیرساختی و اجزاء شبکه موارد: شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات دارایی داده و حاکمیت داده موارد: مجموعه داده‌های تجاری، داده‌های بازار راه‌حل های نوین بانکداری موارد: مدیریت API	جعل هویت دستگاه/کاربر سوءاستفاده از هویت کاربر برای کلاهبرداری	استفاده از روش‌های قوی احراز هویت، ثبت فعالیت‌های کاربران برای پیگیری احراز هویت چندعاملی قوی بیومتریک رفتاری نظارت بر فعالیت‌های مشکوک آموزش کاربران
تزریق اسکریپت‌های بین سایتی	دارایی‌های مالی موارد: ذخایر سرمایه‌ای، ذخایر سرمایه‌ای، سبد سرمایه‌گذاری بانک دارایی‌های زیرساختی و اجزاء شبکه موارد: سامانه‌های نظارتی و انطباق دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های سرمایه‌گذاری، سامانه‌های پردازش تراکنش، سامانه‌های مدیریت منابع انسانی، سامانه‌های حقوقی و انطباق، سامانه‌های گزارش‌دهی و ریسک، پلتفرم‌های وام و اعتبار دارایی نامشهود موارد: روابط با مشتریان و شرکا، مجوزهای قانونی و نظارتی داده و حاکمیت داده موارد: داده‌های مشتری، رضایت و حریم خصوصی مشتری، حاکمیت و مدیریت داده، داده‌های شرکای تجاری، داده‌های طرف سوم، داده‌های کارکنان، مجموعه داده‌های تجاری، داده‌های نظارتی و انطباقی، اطلاعات مشتری، داده‌های تحلیلی و پیش‌بینی، داده‌های بازار، شاخص‌های بازار مالی، دارایی خدمات موارد: خدمات پرداخت، خدمات وام و اعتبار، خدمات سرمایه‌گذاری، خدمات بیمه‌ای مرتبط با بانکداری دارایی‌های امنیت فناوری اطلاعات موارد: سیستم‌های رمزنگاری راه‌حل های نوین بانکداری موارد: مدیریت API سیستم‌های حفاظت از داده و انطباق موارد: حفاظت از داده‌ها، مدیریت انطباق، کنترل دسترسی و حریم خصوصی دارایی‌های عملیاتی موارد: پلتفرم GRCT، حاکمیت، مدیریت ریسک، کنترل‌های داخلی	اجرای اسکریپت‌های مخرب در مرورگر کاربران اجرای کد مخرب در مرورگر کاربر	اعتبارسنجی ورودی‌ها با Whitelisting، محدود کردن دسترسی به پایگاه داده CSP محدودسازی کوکی‌ها (HTTP-only) جلوگیری از اسکریپت خارجی

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تزریق SQL	دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های تحلیل داده و هوش تجاری دارایی داده و حاکمیت داده موارد: شاخص‌های مبتنی بر اینترنت	دسترسی غیرمجاز به پایگاه داده دستکاری یا استخراج اطلاعات از پایگاه داده	استفاده از Prepared Statements، محدود کردن دسترسی به پایگاه داده استفاده از ORM پارامترگذاری WAF بررسی queryهای غیرعادی
جعل عمیق	دارایی داده و حاکمیت داده موارد: داده‌های تحلیلی و پیش‌بینی دارایی‌های امنیت فناوری اطلاعات موارد: حفاظت از نقاط پایانی	جعل هویت از طریق جعل عمیق‌ها جعل واقعیت برای فریب تصمیم‌گیران	احراز هویت چندعاملی، ثبت اطلاعات منشأ محتوا GAN fingerprinting راستی‌آزمایی چندعاملی کلمات عبور گفتاری پویا
بازپخش	دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های مدیریت منابع انسانی دارایی داده و حاکمیت داده موارد: داده‌های مشتری،داده‌های، شرکای‌تجاری، داده‌های طرف سوم، داده‌های کارکنان دارایی نامشهود موارد:مالکیت فکری	سوءاستفاده از داده‌ها/فرمان‌های معتبر اجرای مجدد تراکنش یا پیام تایید شده	استفاده از زمان‌بندی و شماره‌های تصادفی، تأیید هویت دو مرحله‌ای، Tokenهای یک‌بار مصرف، بررسی timestamp، Challenge-response،
تهدیدهای پایدار پیشرفته	دارایی‌های مالی موارد: ذخایر سرمایه‌ای، سرمایه نقدی و سپرده‌ها،سبد سرمایه‌گذاری بانک، دارایی‌های زیرساخت فناوری اطلاعات موارد: سخت‌افزارهای بانکی، نرم‌افزار بانکداری متمرکز،شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات،تجهیزات پردازش تراکنش و سرورهای مالی، سامانه‌های فناوری اطلاعات، سامانه‌های نظارتی و انطباق، سامانه‌های امنیت سایبری، دستگاه‌های خودپرداز، مراکز داده دارایی‌های زیرساخت فناوری اطلاعات موارد: پشتیبان‌گیری و بازیابی، زیرساخت ابری، پلتفرم‌های مدیریت مشتری،پلتفرم‌های سرمایه‌گذاری، سامانه‌های پردازش تراکنش، سامانه‌های مدیریت منابع انسانی، سامانه‌های حقوقی و انطباق، سامانه‌های گزارش‌دهی و ریسک، پلتفرم‌های آموزش و توسعه، پلتفرم‌های وام و اعتبار، پلتفرم‌های تحلیل داده و هوش تجاری دارایی نامشهود موارد: برند و اعتبار بانک، روابط با مشتریان و شرکا.، مجوزهای قانونی و نظارتی، مالکیت فکری	نفوذ بلندمدت به سیستم‌ها/ داده‌ها نفوذ بلندمدت و مخفی در شبکه یا سامانه	ثبت تمامی فعالیت‌ها، راه‌اندازی سیستم‌های بازیابی اضطراری EDR/XDR UEBA تفکیک شبکه + اصل کمترین دسترسی تحلیل تهدیدات (Threat Intelligence) تشخیص ناهنجاری با هوش مصنوعی مدیریت به‌روزرسان وصله‌ها

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تهدیدهای پایدار پیشرفته	داده و حاکمیت داده موارد: داده‌های مشتری،رضایت و حریم خصوصی مشتری، حاکمیت و مدیریت داده، داده‌های شرکای تجاری، داده‌های طرف سوم، داده‌های کارکنان، شاخص‌های مبتنی بر اینترنت، مجموعه داده‌های تجاری، داده‌های نظارتی و انطباقی، اطلاعات مشتری، داده‌های تحلیلی و پیش‌بینی، شاخص‌های بازار مالی دارایی خدمات موارد:خدمات بانکداری دیجیتال، خدمات پرداخت، خدمات وام و اعتبار، خدمات سرمایه‌گذاری، خدمات بیمه‌ای مرتبط با بانکداری دارایی‌های امنیت فناوری اطلاعات موارد: سیستم‌های SIEM، مدیریت آسیب‌پذیری‌ها، پاسخ به حوادث امنیتی، امنیت شبکه، سیستم‌های رمزنگاری، راهکارهای مدیریت هویت و دسترسی (IAM)، حفاظت از نقاط پایانی، سیستم‌های SIEM، مدیریت آسیب‌پذیری‌ها، پاسخ به حوادث امنیتی راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی،کیف پول دیجیتال، بانکداری همراه، راهکارهای نئوبانک، مدیریت API سیستم‌های حفاظت از داده و انطباق موارد: حفاظت از داده‌ها، مدیریت انطباق، کنترل دسترسی و حریم خصوصی، مدیریت حوادث و فارنزیک، پشتیبان‌گیری و بازیابی دارایی‌های عملیاتی موارد: پلتفرم GRC، حاکمیت، مدیریت ریسک، مدیریت حوادث، کنترل‌های داخلی	نفوذ بلندمدت به سیستم‌ها/ داده‌ها نفوذ بلندمدت و مخفی در شبکه یا سامانه	ثبت تمامی فعالیت‌ها، راه‌اندازی سیستم‌های بازیابی اضطراری EDR/XDR UEBA تفکیک شبکه + اصل کمترین دسترسی تحلیل تهدیدات (Threat Intelligence) تشخیص ناهنجاری با هوش مصنوعی مدیریت به‌روزرسان وصله‌ها
توزیع بدافزار	دارایی‌های زیرساختی و اجزاء شبکه موارد: سخت‌افزارهای بانکی، دستگاه‌های خودپرداز دارایی‌های امنیت فناوری اطلاعات موارد: حفاظت از نقاط پایانی	آلوده‌سازی دستگاه‌ها/ خدمات گسترش آلودگی در کل زیرساخت	اعتبارسنجی داده‌ها با هش، فیلتر کردن ترافیک شبکه بررسی ترافیک خروجی EDR مدیریت دسترسی

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
بدافزارهای خودآموز	دارایی‌های زیرساختی و اجزاء شبکه موارد: شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات، سامانه‌های فناوری اطلاعات دارایی‌های امنیت فناوری اطلاعات موارد: حفاظت از نقاط پایانی	فرار از سیستم‌های تشخیص با استفاده از یادگیری ماشین فرار از آنتی‌ویروس و تسخیر سیستم	استفاده از سیستم‌های تشخیص نفوذ، به‌روزرسانی منظم نرم‌افزارها استفاده از فناوری‌های تشخیص مبتنی بر یادگیری ماشین (AI-based malware detection) جداسازی و sandboxing برنامه‌ها و فایل‌های مشکوک به‌روزرسانی مستمر سیستم‌های تشخیص تهدیدات (Endpoint Detection & Response - EDR) استفاده از مدل‌های تشخیص ناهنجاری برای رفتارهای مشکوک در شبکه و Endpoint
کلون‌سازی	دارایی‌های امنیت فناوری اطلاعات موارد: راهکارهای مدیریت هویت و دسترسی (IAM) راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی	تکثیر غیرمجاز دستگاه‌ها/ خدمات استخراج داده از طریق نشت فیزیکی یا زمانی	احراز هویت بیومتریک، ثبت فعالیت‌های کاربران برای ردیابی تشخیص الگو استخراج رمزنگاری و مبهم‌سازی مدل Behavioral Biometrics
مهندسی معکوس	دارایی‌های مالی موارد: دارایی‌های زیرساختی و اجزاء شبکه مراکز داده دارایی‌های زیرساخت فناوری اطلاعات موارد: سرورها، دستگاه‌های شبکه، راهکارهای ذخیره‌سازی، زیرساخت ابری دارایی نامشهود موارد:روابط با مشتریان و شرکا. دارایی داده و حاکمیت داده موارد:داده‌های نظارتی و انطباقی دارایی خدمات موارد: خدمات پرداخت، خدمات بانکداری دیجیتال دارایی‌های امنیت فناوری اطلاعات موارد: سیستم‌های رمزنگاری، راهکارهای مدیریت هویت و دسترسی (IAM)، سیستم‌های SIEM، پاسخ به حوادث امنیتی راه‌حل های نوین بانکداری موارد: کیف پول دیجیتال سیستم‌های حفاظت از داده و انطباق موارد: مدیریت انطباق،کنترل دسترسی و حریم خصوصی،مدیریت حوادث و فارنزیک،پشتیبان‌گیری و بازیابی دارایی‌های عملیاتی موارد: مدیریت حوادث	افشای ساختار سیستم و آسیب‌پذیری‌ها	کد Obfuscation سخت‌سازی باینری لایسنس سخت‌افزاری مدل Watermarking

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
کانال جانبی	دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های مدیریت منابع انسانی دارایی نامشهود موارد: روابط با مشتریان و شرکا.	استنتاج غیرمجاز داده‌های حساس استخراج داده از طریق نشت فیزیکی یا زمانی	رمزنگاری داده‌ها، محدود کردن دسترسی به داده‌های حساس، اعتبارسنجی داده‌ها،
بدافزارهای مبتنی بر هوش مصنوعی برای فرار از تشخیص	دارایی‌های مالی موارد: شاخص‌های بازار مالی	فرار از سیستم‌های امنیتی مبتنی بر هوش مصنوعی فرار از مکانیزم‌های دفاعی سنتی	استفاده از هوش مصنوعی برای شناسایی بدافزارها، به‌روزرسانی منظم سیستم‌ها تحلیل رفتاری دینامیک sandbox هوشمند ترکیب signature + anomaly detection
دور زدن سامانه‌های تشخیص تهدیدات داخلی	دارایی‌های زیرساختی و اجزاء شبکه موارد: سامانه‌های نظارتی و انطباق دارایی‌های امنیت فناوری اطلاعات موارد: مدیریت آسیب‌پذیری‌ها سیستم‌های حفاظت از داده و انطباق موارد: مدیریت حوادث و فارنزیک	عدم تشخیص تهدیدات داخلی نفوذ پنهانی بدون هشدار	استفاده از سیستم‌های چندلایه برای تشخیص، فیلتر کردن ترافیک غیرمعمول IDS/IPS پیشرفته، همبست‌سازی لاگ‌ها
دور زدن فیلترهای ایمیل	دارایی‌های زیرساختی و اجزاء شبکه موارد: سامانه‌های امنیت سایبری داده و حاکمیت داده موارد: داده (داده‌های ارتباطی) دارایی‌های امنیت فناوری اطلاعات موارد: امنیت شبکه، راهکارهای مدیریت هویت و دسترسی (IAM)	توزیع محتوای مخرب (مثل فیشینگ) اجرای موفق حملات فیشینگ	فیلتر کردن ترافیک خروجی، اعتبارسنجی ارسال‌کنندگان با SPF/DKIM SPF/DKIM، فیلتر هوشمند، آموزش کاربران
فرار از سامانه تشخیص نفوذ شبکه	دارایی‌های زیرساختی و اجزاء شبکه موارد: سامانه‌های نظارتی و انطباق دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های مدیریت منابع انسانی، سامانه‌های حقوقی و انطباق، پلتفرم‌های وام و اعتبار دارایی داده و حاکمیت داده موارد: داده‌های مشتری، رضایت و حریم خصوصی مشتری، داده‌های شرکای تجاری، داده‌های طرف سوم، داده‌های نظارتی و انطباقی،اطلاعات مشتری، داده‌های تحلیلی و پیش‌بینی، دارایی خدمات، موارد: خدمات بانکداری دیجیتال، سیستم‌های حفاظت از داده و انطباق موارد: حفاظت از داده‌ها، مدیریت انطباق، کنترل دسترسی و حریم خصوصی	نفوذ بدون شناسایی نفوذ پنهان از دید IDS/IPS شبکه	استفاده از سیستم‌های چندلایه برای تشخیص، فیلتر کردن ترافیک غیرمعمول رفتارمحور SIEM + EDR+ بروزرسانی مرتب ruleها تحلیل ترافیک رمزنگاری‌شده ترکیب signature و anomaly detection

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
سازگارسازی حمله	دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های تحلیل داده و هوش تجاری دارایی داده و حاکمیت داده موارد: داده‌های تحلیلی و پیش‌بینی دارایی‌های عملیاتی موارد: دارایی‌های غیرنقدی	فرار از سیستم‌های تشخیص با تغییر رفتار تنظیم حمله براساس رفتار سیستم هدف	آنالیز ترافیک برای شناسایی الگوهای جدید، فیلتر کردن ترافیک مشکوک دفاع تطبیقی تنوع در راهکارها Red/Blue Team به‌روزرسانی Intelligence و امضاها
ثبت کلید ضمنی	دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های گزارش‌دهی و ریسک، حملات شناسایی شده مربوط به دارایی‌های نامشهود دارایی داده و حاکمیت داده موارد: شاخص‌های بازار مالی، دارایی خدمات خدمات بانکداری دیجیتال دارایی‌های عملیاتی موارد: مدیریت ریسک	سرقت کلیدهای رمزنگاری از طریق نشت اطلاعات جانبی ضبط کلیدهای فشرده‌شده توسط مهاجم	ثبت فعالیت‌ها با زمان‌بندی دقیق، رمزنگاری کلیدها در حین انتقال استفاده از کیبورد مجازی تحلیل رفتار نرم‌افزارها Anti-keylogger Tools
سرقت اعتبارنامه	دارایی‌های مالی موارد: ذخایر سرمایه‌ای، سرمایه نقدی و سپرده‌ها دارایی‌های زیرساختی و اجزاء شبکه موارد:سخت‌افزارهای بانکی،نرم‌افزار بانکداری متمرکز،تجهیزات پردازش تراکنش و سرورهای مالی، سامانه‌های فناوری اطلاعات، سامانه‌های نظارتی و انطباق، سامانه‌های امنیت سایبری، دستگاه‌های خودپرداز، مراکز داده دارایی‌های زیرساخت فناوری اطلاعات موارد: راهکارهای ذخیره‌سازی، پشتیبان‌گیری و بازیابی، زیرساخت ابریپلتفرم‌های مدیریت مشتری پلتفرم‌های سرمایه‌گذاری، سامانه‌های پردازش تراکنش، سامانه‌های مدیریت منابع انسانی، سامانه‌های حقوقی و انطباق، سامانه‌های گزارش‌دهی و ریسک، پلتفرم‌های آموزش و توسعه،پلتفرم‌های وام و اعتبار، پلتفرم‌های تحلیل داده و هوش تجاری دارایی داده و حاکمیت داده موارد: داده‌های بازار، شاخص‌های بازار مالی دارایی خدمات موارد: خدمات بانکداری دیجیتال، خدمات پرداخت، خدمات وام و اعتبار، خدمات سرمایه‌گذاری، خدمات بیمه‌ای مرتبط با بانکداری، دارایی‌های امنیت فناوری اطلاعات موارد: سیستم‌های رمزنگاری، سیستم‌های SIEM، پاسخ به حوادث امنیتی	دسترسی غیرمجاز به سیستم‌های مدیریتی دسترسی کامل به منابع یا جعل هویت	رمزنگاری اعتبارنامه‌ها در حین انتقال، محدود کردن دسترسی به سیستم‌های حساس احراز هویت چندعاملی هشینگ bcrypt بررسی نشت از طریق threat-feeds لاگین‌های مشکوک Lockout حساب

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
سرقت اعتبارنامه	راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی، راهکارهای نئوبانک، مدیریت API سیستم‌های حفاظت از داده و انطباق موارد: حفاظت از داده‌ها، مدیریت انطباق، کنترل دسترسی و حریم خصوصی، مدیریت حوادث و فارنزیک، پشتیبان‌گیری و بازیابی دارایی‌های عملیاتی موارد: پلتفرم GRC، حاکمیت، مدیریت ریسک، مدیریت حوادث، کنترل‌های داخلی	دسترسی غیرمجاز به سیستم‌های مدیریتی دسترسی کامل به منابع یا جعل هویت	رمزنگاری اعتبارنامه‌ها در حین انتقال، محدود کردن دسترسی به سیستم‌های حساس احراز هویت چندعاملی هشینگ bcrypt بررسی نشت از طریق threat-feeds لاگین‌های مشکوک Lockout حساب
یافتن مسیر	دارایی‌های زیرساختی و اجزاء شبکه موارد: تجهیزات پردازش تراکنش و سرورهای مالی، مراکز داده دارایی‌های زیرساخت فناوری اطلاعات موارد: مراکز داده دارایی داده و حاکمیت داده موارد: طلاعات مشتری، داده‌های تحلیلی و پیش‌بینی دارایی خدمات موارد: خدمات بیمه‌ای مرتبط با بانکداری سیستم‌های حفاظت از داده و انطباق موارد: پشتیبان‌گیری و بازیابی	نقشه‌برداری آسیب‌پذیری‌های شبکه استخراج توپولوژی شبکه برای حملات بعدی	استفاده از پروتکل‌های امن برای مسیریابی، اعتبارسنجی داده‌های مسیریابی رمزنگاری مسیرها بررسی گره‌های میانجی مقاوم‌سازی پروتکل‌های مسیریابی
حرکت جانبی	دارایی‌های زیرساختی و اجزاء شبکه موارد: سخت‌افزارهای بانکی، تجهیزات پردازش تراکنش و سرورهای مالی، تجهیزات ذخیره‌سازی، سامانه‌های فناوری اطلاعات، دستگاه‌های خودپرداز دارایی‌های زیرساخت فناوری اطلاعات موارد: سرورها دارایی خدمات موارد: خدمات بانکداری دیجیتال، خدمات پرداخت دارایی‌های امنیت فناوری اطلاعات موارد: سیستم‌های رمزنگاری، حفاظت از نقاط پایانی راه‌حل های نوین بانکداری موارد: بانکداری همراه، کیف پول دیجیتال	گسترش حملات در شبکه گسترش نفوذ مهاجم در شبک	استفاده از RBAC برای کنترل دسترسی، آنالیز ترافیک شبکه برای شناسایی فعالیت‌های غیرطبیعی Segment کردن شبکه تحلیل ترافیک داخلی SIEM Honeynet

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
استخراج داده‌ها	دارایی‌های زیرساختی و اجزاء شبکه موارد: نرم‌افزار بانکداری متمرکز، تجهیزات پردازش تراکنش و سرورهای مالی، سامانه‌های فناوری اطلاعات، سامانه‌های نظارتی و انطباق، دستگاه‌های خودپرداز دارایی‌های زیرساخت فناوری اطلاعات موارد: سرورها، راهکارهای ذخیره‌سازی، پشتیبان‌گیری و بازیابی، پلتفرم‌های سرمایه‌گذاری، سامانه‌های پردازش تراکنش، پلتفرم‌های وام و اعتبار دارایی داده و حاکمیت داده موارد: مجموعه داده‌های تجاری، طلاعات مشتری، داده‌های بازار، شاخص‌های بازار مالی دارایی خدمات موارد: خدمات بانکداری دیجیتال، خدمات پرداخت، خدمات وام و اعتبار، خدمات سرمایه‌گذاری، خدمات بیمه‌ای مرتبط با بانکداری دارایی‌های امنیت فناوری اطلاعات موارد: امنیت شبکه، سیستم‌های رمزنگاری، سیستم‌های SIEM، مدیریت آسیب‌پذیری‌ها، پاسخ به حوادث امنیتی راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی، بانکداری همراه، کیف پول دیجیتال، راهکارهای نئوبانک، مدیریت API، سیستم‌های حفاظت از داده و انطباق موارد: پشتیبان‌گیری و بازیابی دارایی‌های عملیاتی موارد: مدیریت حوادث	انتقال غیرمجاز داده‌های حساس نشت اطلاعات محرمانه به خارج سازمان	استفاده از امضای دیجیتال برای ردیابی داده‌ها، رمزنگاری داده‌های حساس DLP هشدار بلادرنگ خروج داده کنترل USB و دسترسی UEBA
تغییر فیلم‌های نظارتی	دارایی‌های زیرساختی و اجزاء شبکه موارد: شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های حقوقی و انطباق دارایی‌های امنیت فناوری اطلاعات موارد: حفاظت از نقاط پایانی	پنهان‌سازی فعالیت‌های مخرب حذف یا تحریف شواهد تصویری	استفاده از امضای دیجیتال برای تأیید صحت ویدیو، ثبت تغییرات با زمان‌بندی واترمارک، امضای دیجیتال، ذخیره امن رمزگذاری امن ویدئوها با رمزنگاری انتها به انتها (End-to-End Encryption) ثبت امضای دیجیتال روی ویدئوها برای اطمینان از اصالت آن‌ها استفاده از سامانه‌های تحلیل ویدیویی برای شناسایی دستکاری‌ها (Video Forensics) ذخیره‌سازی نسخه‌های پشتیبان و نظارت مداوم بر یکپارچگی ویدئوها

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
حمله حریم خصوصی	دارایی‌های مالی موارد: ذخایر سرمایه‌ای، سرمایه نقدی و سپرده‌ها،دارایی‌های غیرنقدی،سبد سرمایه‌گذاری بانک، دارایی‌های زیرساختی و اجزاء شبکه دارایی‌های زیرساخت فناوری اطلاعات موارد: سخت‌افزارهای بانکی، نرم‌افزار بانکداری متمرکز،شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات،تجهیزات پردازش تراکنش و سرورهای مالی،تجهیزات ذخیره‌سازی، سامانه‌های فناوری اطلاعات، سامانه‌های نظارتی و انطباق، سامانه‌های امنیت سایبری، دستگاه‌های خودپرداز، مراکز داده دارایی‌های زیرساخت فناوری اطلاعات موارد: پشتیبان‌گیری و بازیابی، زیرساخت ابری، پلتفرم‌های مدیریت مشتری،پلتفرم‌های سرمایه‌گذاری، سامانه‌های پردازش تراکنش، سامانه‌های مدیریت منابع انسانی، سامانه‌های حقوقی و انطباق، سامانه‌های گزارش‌دهی و ریسک، پلتفرم‌های آموزش و توسعه، پلتفرم‌های وام و اعتبار، پلتفرم‌های تحلیل داده و هوش تجاری دارایی نامشهود موارد: روابط با مشتریان و شرکا،، مجوزهای قانونی و نظارتی داده و حاکمیت داده موارد:رضایت و حریم خصوصی مشتری، حاکمیت و مدیریت داده،داده‌های شرکای تجاری، داده‌های طرف سوم، داده‌های کارکنان، مجموعه داده‌های تجاری، داده‌های نظارتی و انطباقی، اطلاعات مشتری، دارایی خدمات موارد:خدمات بانکداری دیجیتال، خدمات پرداخت، خدمات وام و اعتبار، خدمات سرمایه‌گذاری، خدمات بیمه‌ای مرتبط با بانکداری دارایی‌های امنیت فناوری اطلاعات موارد: سیستم‌های SIEM، مدیریت آسیب‌پذیری‌ها،طط پاسخ به حوادث امنیتی راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی، بانکداری همراه، راهکارهای نئوبانک، مدیریت API سیستم‌های حفاظت از داده و انطباق	لو رفتن داده‌های شخصی کاربران افشای اطلاعات شخصی و قانونی	رمزنگاری داده‌های شخصی، کنترل دسترسی به پرونده‌های مشتریان Differential Privacy بررسی نشت اطلاعات (Membership/Model Inference) رمزنگاری پیش‌بینی/آموزش (Homomorphic Encryption) RBAC روی داده‌های حساس

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
حمله حریم خصوصی	موارد: حفاظت از داده‌ها، مدیریت انطباق، کنترل دسترسی و حریم خصوصی، مدیریت حوادث و فارنزیک، پشتیبان‌گیری و بازیابی دارایی‌های عملیاتی موارد: پلتفرم GRC، حاکمیت، مدیریت ریسک، مدیریت حوادث، کنترل‌های داخلی	لو رفتن داده‌های شخصی کاربران افشای اطلاعات شخصی و قانونی	رمزنگاری داده‌های شخصی، کنترل دسترسی به پرونده‌های مشتریان Differential Privacy بررسی نشت اطلاعات (Membership/Model Inference) رمزنگاری پیش‌بینی/آموزش (Homomorphic Encryption) RBAC روی داده‌های حساس
هماهنگی حملات	دارایی‌های زیرساختی و اجزاء شبکه موارد: تجهیزات پردازش تراکنش و سرورهای مالی، تجهیزات ذخیره‌سازی دارایی‌های زیرساخت فناوری اطلاعات موارد: سرورها، دستگاه‌های شبکه، راهکارهای ذخیره‌سازی، پشتیبان‌گیری و بازیابی دارایی داده و حاکمیت داده موارد: مجموعه داده‌های تجاری، داده‌های بازار سیستم‌های حفاظت از داده و انطباق موارد: پشتیبان‌گیری و بازیابی	حملات هماهنگ و گسترده شکست همزمان چند لایه دفاعی	ثبت فعالیت‌های مشکوک، فیلتر کردن ترافیک هماهنگ تحلیل هم‌زمان ترافیک و لاگ‌ها (SIEM) سیاست‌های سخت‌گیرانه بین‌سازمانی Honeypot و سیستم‌های فریب SOAR تیم CSIRT
سیاه چاله	دارایی‌های مالی موارد: ذخایر سرمایه‌ای دارایی‌های زیرساختی و اجزاء شبکه موارد: سخت‌افزارهای بانکی، نرم‌افزار بانکداری متمرکز، تجهیزات پردازش تراکنش و سرورهای مالی، امانه‌های فناوری اطلاعات، دستگاه‌های خودپرداز دارایی‌های زیرساخت فناوری اطلاعات موارد: سرورها، راهکارهای ذخیره‌سازی، پشتیبان‌گیری و بازیابی، زیرساخت ابری، پلتفرم‌های سرمایه‌گذاری، سامانه‌های پردازش تراکنش دارایی نامشهود موارد: مجوزهای قانونی و نظارتی دارایی داده و حاکمیت داده موارد: داده‌های مشتری، داده‌های شرکای تجاری، داده‌های طرف سوم، داده‌های کارکنان دارایی خدمات موارد: خدمات بانکداری دیجیتال، خدمات پرداخت، خدمات بیمه‌ای مرتبط با بانکداری دارایی‌های امنیت فناوری اطلاعات موارد: راهکارهای مدیریت هویت و دسترسی (IAM) راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی،بانکداری همراه،کیف پول دیجیتال،راهکارهای نئوبانک سیستم‌های حفاظت از داده و انطباق موارد:کنترل دسترسی و حریم خصوصی	اختلال در مسیریابی داده‌ها قطع یا جذب ترافیک شبکه	مراقبت از دسترسی به گره‌های کلیدی، اعتبارسنجی داده‌های ورودی مسیریابی مقاوم بررسی چند مسیر هم‌زمان پروتکل AODV-SEC تشخیص گره مخرب

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
دستکاری رکوردها	دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های گزارش‌دهی و ریسک دارایی داده و حاکمیت داده موارد: مجموعه داده‌های تجاری سیستم‌های حفاظت از داده و انطباق موارد: پشتیبان‌گیری و بازیابی	تغییر غیرمجاز داده‌های ذخیره‌شده تحریف داده‌های حساس یا مالی	استفاده از چک‌سازی دیجیتال، ثبت تمامی تغییرات با زمان‌بندی کنترل نسخه، هش داده، نظارت لحظه‌ای پیاده‌سازی سیستم‌های مدیریت هویت و دسترسی (IAM) ثبت رویدادهای امنیتی (Logging) و نگهداری نسخه‌های پشتیبان (Backups) مانیتورینگ و سیستم‌های تشخیص دسترسی‌های غیرمجاز Intrusion Detection Systems - (IDS) رمزنگاری و امضای دیجیتال سوابق حساس
حمله DDoS	دارایی‌های مالی موارد:سرمایه نقدی و سپرده‌ها دارایی‌های زیرساخت فناوری اطلاعات موارد:سامانه‌های مدیریت منابع انسانی دارایی نامشهود موارد:روابط با مشتریان و شرکا. دارایی داده و حاکمیت داده موارد: داده‌های مشتری، رضایت و حریم خصوصی مشتری، داده‌های شرکای تجاری، داده‌های طرف سوم، داده‌های کارکنان، اطلاعات مشتری سیستم‌های حفاظت از داده و انطباق موارد: کنترل دسترسی و حریم خصوصی	اختلال در دسترس‌پذیری خدمات/شبکه قطع دسترسی کاربران به سرویس	استفاده از CDN و فیلتر کردن ترافیک، محدود کردن دسترسی به سرویس‌ها CDN با Anycast Load Balancing سرویس mitigation حرفه‌ای
باج افزارهای مبتنی بر هوش مصنوعی	دارایی‌های زیرساختی و اجزاء شبکه موارد: شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات، سامانه‌های فناوری اطلاعات دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های گزارش‌دهی و ریسک	رمزگذاری یا حذف داده‌های حیاتی رمزگذاری هوشمند و گریز از دفاع	ثبت فعالیت‌های کاربران، راه‌اندازی سیستم‌های بازیابی اضطراری Backup امن و آفلاین EDR/XDR کنترل دسترسی رفتارشناسی فایل‌ها تحلیل تهدیدات

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
جمینگ	دارایی‌های زیرساختی و اجزاء شبکه موارد: شبکه‌های ارتباطی امن و امنیت فناوری اطلاعات،سامانه‌های فناوری اطلاعات،سامانه‌های امنیت سایبری، دستگاه‌های خودپرداز، مراکز داده دارایی‌های زیرساخت فناوری اطلاعات موارد: سرورها دارایی داده و حاکمیت داده موارد: داده‌های بازار دارایی‌های امنیت فناوری اطلاعات موارد: امنیت شبکه، مدیریت آسیب‌پذیری‌ها	اختلال در ارتباطات بی‌سیم قطع ارتباط رادیویی یا بی‌سیم	طیف فرکانسی تطبیقی تشخیص سیگنال‌های مزاحم استفاده از فرکانس‌های متعدد

جدول ۶: حملات هوش مصنوعی تخاصمی در حوزه مالی و بانکی

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
استخراج مدل	دارایی‌های امنیت فناوری اطلاعات موارد: راهکارهای مدیریت هویت و دسترسی (IAM) راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی	سرقت مدل های یادگیری ماشین اختصاصی. کپی غیرمجاز از مدل‌های یادگیری ماشین	مبهم‌سازی ساختار مدل، محدود کردن پرس‌وجوهای تکراری به API محدود کردن دسترسی به مدل محرمانه‌سازی خروجی‌ها تشخیص و مسدودسازی رفتار غیرعادی استفاده از مدل‌های جایگزین و پراکسی
بازسازی داده	دارایی‌های زیرساختی و اجزاء شبکه موارد: نرم‌افزار بانکداری متمرکز، سامانه‌های فناوری اطلاعات دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های گزارش‌دهی و ریسک، پلتفرم‌های آموزش و توسعه، پلتفرم‌های تحلیل داده و هوش تجاری دارایی داده و حاکمیت داده موارد: شاخص‌های مبتنی بر اینترنت، داده‌های تحلیلی و پیش‌بینی، داده‌های بازار، شاخص‌های بازار مالی دارایی خدمات موارد: خدمات بانکداری دیجیتال دارایی‌های امنیت فناوری اطلاعات موارد: پاسخ به حوادث امنیتی، سیستم‌های SIEM، حفاظت از نقاط پایانی، امنیت شبکه راه‌حل های نوین بانکداری موارد: بانکداری همراه، راهکارهای نئوبانک، مدیریت API سیستم‌های حفاظت از داده و انطباق موارد: حفاظت از داده‌ها، مدیریت حوادث و فارنزیک دارایی‌های عملیاتی موارد: مدیریت ریسک	بازیابی داده‌های حساس از قطعات بازیابی داده اصلی از خروجی مدل	استفاده از تکه‌سازی امن داده‌ها ، احراز هویت برای فرآیند بازسازی Secure Erase (SSD) پاک‌سازی چندمرحله‌ای نویز کنترل‌شده کنترل دسترسی رسانه‌ها

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
استنتاج عضویت	دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های گزارش‌دهی و ریسک، پلتفرم‌های تحلیل داده و هوش تجاری دارایی داده و حاکمیت داده موارد: داده‌های تحلیلی و پیش‌بینی، شاخص‌های بازار مالی دارایی خدمات موارد: خدمات وام و اعتبار دارایی‌های عملیاتی موارد: پلتفرم GRC	لو رفتن داده‌های مورد استفاده در مدل‌های هوش مصنوعی شناسایی اینکه داده‌ای در آموزش مدل بوده یا نه	افزودن نویز کنترل‌شده به داده‌ها، محدود کردن دسترسی به خروجی مدل استفاده از Differential Privacy ارزیابی ریسک پاسخ مدل بررسی درز اطلاعات آموزشی
استنتاج ویژگی	دارایی‌های مالی موارد: سبد سرمایه‌گذاری بانک دارایی‌های زیرساختی و اجزاء شبکه موارد: تجهیزات پردازش تراکنش و سرورهای مالی دارایی‌های زیرساخت فناوری اطلاعات موارد: سرورها، سامانه‌های گزارش‌دهی و ریسک، پلتفرم‌های تحلیل داده و هوش تجاری دارایی‌های عملیاتی موارد: مدیریت ریسک	لو رفتن ویژگی‌های پنهان داده‌ها استخراج ویژگی‌های خصوصی از مدل	استفاده از رمزنگاری ویژگی‌ها، نظارت بر پرس‌وجوهای غیرمعمول به مدل Differential Privacy محدودسازی خروجی مدل بررسی ریسک مدل
مسمومیت مدل هوش مصنوعی	دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های گزارش‌دهی و ریسک دارایی داده و حاکمیت داده موارد: داده‌های تحلیلی و پیش‌بینی، شاخص‌های مبتنی بر اینترنت	مسموم سازی خروجی‌های مدل‌های هوش مصنوعی کاهش دقت و قابلیت اعتماد مدل	اعتبارسنجی یکپارچگی داده‌های آموزشی، استفاده از الگوریتم‌های تشخیص داده‌های مسموم استفاده از داده معتبر Data sanitization مانیتور Drift داده مدل‌های Robust ML Adversarial Training اعتبارسنجی داده آموزشی محدودسازی دسترسی به مدل
مسمومیت با برچسب تمیز	دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های آموزش و توسعه دارایی نامشهود موارد: روابط با مشتریان و شرکا. دارایی داده و حاکمیت داده موارد: داده‌های کارکنان، اطلاعات مشتری، داده‌های بازار، شاخص‌های بازار مالی دارایی خدمات موارد: خدمات بانکداری دیجیتال، خدمات پرداخت، خدمات وام و اعتبار دارایی‌های امنیت فناوری اطلاعات موارد: راهکارهای مدیریت هویت و دسترسی (IAM)	مسموم‌سازی ظریف مدل‌های هوش مصنوعی آموزش مدل با داده‌های برچسب‌خورده اشتباه	نظارت بر انحرافات خروجی مدل، به‌روزرسانی مدل با داده‌های معتبر بررسی تعادل کلاس‌ها consistency checking پایش Model Drift

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
مسمومیت با برچسب تمیز	راه‌حل های نوین بانکداری موارد: کیف پول دیجیتال، بانکداری اینترنتی، سیستم‌های حفاظت از داده و انطباق موارد: کنترل دسترسی و حریم خصوصی	مسموم‌سازی ظریف مدل‌های هوش مصنوعی آموزش مدل با داده‌های برچسب‌خورده اشتباه	نظارت بر انحرافات خروجی مدل، به‌روزرسانی مدل با داده‌های معتبر بررسی تعادل کلاس‌ها consistency checking پایش Model Drift
مسمومیت هدفمند	دارایی داده و حاکمیت داده موارد: داده‌های تحلیلی و پیش‌بینی، شاخص‌های مبتنی بر اینترنت دارایی‌های امنیت فناوری اطلاعات موارد: مدیریت آسیب‌پذیری‌ها	دستکاری هدفمند خروجی مدل‌ها هدف‌گیری مدل خاص برای تصمیم نادرست	استفاده از مکانیزم‌های دفاعی مبتنی بر افزونگی، آموزش مدل با سناریوهای حمله کنترل دسترسی به داده‌های آموزشی مدل‌های بانکی اعتبارسنجی مستمر و کنترل کیفیت داده‌های آموزشی مانیتورینگ عملکرد مدل‌ها برای شناسایی سریع تغییرات غیرعادی استفاده از الگوریتم‌های مقاوم در برابر داده‌های مسموم (Robust Training Algorithms)
مسمومیت درب پشتی	دارایی‌های زیرساختی و اجزاء شبکه موارد: سامانه‌های فناوری اطلاعات دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های آموزش و توسعه دارایی داده و حاکمیت داده موارد:داده‌های تحلیلی و پیش‌بینی	قرار دادن درگاه‌های پنهان در مدل‌ها قرار دادن رفتار مخرب قابل فعال‌سازی	اسکن مدل برای الگوهای غیرعادی، اعتبارسنجی خروجی در محیط‌های ایزوله ارزیابی امنیتی دوره‌ای مدل‌ها و منابع داده‌ای استفاده از روش‌های دفاعی مانند pruning، fine-tuning و detoxification برای حذف درهای پشتی اعتبارسنجی دقیق داده‌ها و مدل‌ها توسط متخصصین امنیتی
مسمومیت در دسترس پذیری	سیستم‌های حفاظت از داده و انطباق موارد: حفاظت از داده‌ها دارایی‌های امنیت فناوری اطلاعات موارد: حفاظت از نقاط پایانی دارایی داده و حاکمیت داده موارد: شاخص‌های مبتنی بر اینترنت	کاهش دسترس‌پذیری خدمات ناکارآمدسازی کلی مدل یادگیری ماشین	استفاده از تکه‌تکه‌سازی امن داده‌ها ، احراز هویت برای فرآیند بازسازی مانیتورینگ پیوسته برای شناسایی کاهش عمدی دسترس‌پذیری خدمات توزیع و تکثیر سامانه‌ها برای مقابله با نقاط شکست (Redundancy) استفاده از الگوریتم‌های دفاعی در سطح داده‌ها برای شناسایی داده‌های مخرب مدیریت دسترسی و محدودیت نرخ درخواست‌ها (Rate Limiting)

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
گریز یا نمونه های متخاصم	دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های آموزش و توسعه، پلتفرم‌های تحلیل داده و هوش تجاری دارایی داده و حاکمیت داده موارد: شاخص‌های مبتنی بر اینترنت، داده‌های تحلیلی و پیش‌بینی، داده‌های بازار، شاخص‌های بازار مالی دارایی‌های عملیاتی موارد: مدیریت ریسک	سیستم‌های هوش مصنوعی/ یادگیری ماشین گمراه کننده فریب مدل با داده شبه‌واقعی	افزودن نویز کنترل‌شده به داده‌ها، محدود کردن دسترسی به خروجی مدل Detect-and-reject فیلتر پیش‌پردازش Ensemble مدل
حمله مبتنی بر RAG	دارایی‌های زیرساخت فناوری اطلاعات موارد: پلتفرم‌های آموزش و توسعه دارایی‌های خدمات موارد: شاخص‌های مبتنی بر اینترنت، داده‌های تحلیلی و پیش‌بینی	دستکاری محتوای تولید شده توسط هوش مصنوعی سوءاستفاده از ترکیب حافظه+مدل زبانی	اعتبارسنجی منابع داده‌های دانش، نظارت بر خروجی‌های غیرمنطقی محدودسازی دسترسی به داده‌های حساس و محرمانه بانک‌ها ارزیابی مستمر مدل‌های هوش مصنوعی از نظر امنیت و نشت داده کنترل و نظارت بر داده‌های آموزشی و مستندات مورد استفاده توسط مدل‌ها پیاده‌سازی مکانیزم‌های شناسایی و هشدار سریع نشت اطلاعات (Data Leakage Detection)
تزریق مستقیم درخواست	دارایی‌های زیرساخت فناوری اطلاعات موارد: سامانه‌های گزارش‌دهی و ریسک، پلتفرم‌های وام و اعتبار، پلتفرم‌های تحلیل داده و هوش تجاری دارایی داده و حاکمیت داده موارد: داده‌های تحلیلی و پیش‌بینی دارایی خدمات موارد: خدمات وام و اعتبار دارایی‌های عملیاتی موارد: مدیریت ریسک	اجرای فرمان‌های غیرمجاز اجرای دستورات در سرور با ورودی کاربر	اعتبارسنجی ورودی‌ها با Whitelisting، محدود کردن دسترسی بر اساس اصول حداقل امتیاز
تزریق غیرمستقیم درخواست	دارایی نامشهود موارد: روابط با مشتریان و شرکا. دارایی خدمات موارد: خدمات وام و اعتبار راه‌حل های نوین بانکداری موارد: بانکداری اینترنتی	دستکاری غیرمستقیم سیستم‌ها بهره‌برداری از کاربر برای تزریق درخواست	پیاده‌سازی CSRF Tokens، نظارت بر تراکنش‌های زنجیره‌ای ORM + پارامترگذاری کوئری WAF نظارت بر کوئری‌های غیرعادی

۴-۳ شناسایی و مقابله با حملات مبتنی بر هوش مصنوعی در حوزه سلامت

جدول ۷: حملات هوش مصنوعی تهاجمی در حوزه سلامت

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تحلیل ترافیک شبکه	شبکه‌های ارتباطی	نشت داده‌ها, نقض حریم خصوصی	رمزنگاری پروتکل‌ها, تشخیص ناهنجاری
تشخیص آسیب‌پذیری	داده‌های امنیتی و آمار فعالیت‌های روزانه	نقض امنیت, سوءاستفاده از آسیب‌پذیری	آزمایش نفوذ, به‌روزرسانی امنیتی
جاسوسی	شبکه‌های ارتباطی	نشت داده‌ها, نقض حریم خصوصی	رمزنگاری پروتکل‌ها, نظارت بر ترافیک
تحلیل رفتاری برای یافتن راه‌های جدید سوءاستفاده	داده‌های امنیتی و آمار فعالیت‌های روزانه	دسترسی غیرمجاز, بهره‌برداری سیستم	تشخیص ناهنجاری, نظارت مستمر
پیش‌بینی نتایج حملات	داده‌های امنیتی و آمار فعالیت‌های روزانه	کاهش دقت سیستم, خطای تصمیم‌گیری	اعتبارسنجی داده ورودی, ممیزی امنیتی
کش‌کاوی	داده‌های امنیتی و آمار فعالیت‌های روزانه	نقض حریم خصوصی, سوءاستفاده داده	فیلتر داده‌ها, نظارت بر دسترسی
مرد میانی	شبکه‌های ارتباطی	نشت داده‌ها, دسترسی غیرمجاز	رمزنگاری قوی, احراز هویت
تولید خودکار اطلاعات نادرست	تیم‌های نظارتی و انطباق	انتشار اطلاعات نادرست, کاهش اعتماد	ابزارهای تشخیص جعل عمیق, آموزش امنیتی
فیشینگ شخصی سازی‌شده	تیم‌های نظارتی و انطباق	سرقت اطلاعات, دسترسی غیرمجاز	آموزش امنیت سایبری, فیلتر ایمیل
سرقت هویت	تیم‌های نظارتی و انطباق	دسترسی غیرمجاز, نقض حریم خصوصی	احراز هویت چندمرحله‌ای, نظارت بر رفتار
تهدیدهای پایدار پیشرفته	داده‌های امنیتی و آمار فعالیت‌های روزانه	دسترسی غیرمجاز, نشت داده‌ها	نظارت مستمر, کنترل دسترسی
بدافزارهای مبتنی بر هوش مصنوعی برای فرار از تشخیص	شبکه‌های ارتباطی	دسترسی غیرمجاز, نقض امنیت	پایش ترافیک, پچ امنیتی
فرار از سامانه تشخیص نفوذ شبکه	شبکه‌های ارتباطی	کاهش دقت تشخیص, نقض امنیت	به‌روزرسانی الگوریتم‌ها, نظارت مستمر
یافتن مسیر	شبکه‌های ارتباطی	نشت اطلاعات شبکه, نقض امنیت	پایش ترافیک, رمزنگاری

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
حمله حریم خصوصی	داده‌های امنیتی و آمار فعالیت‌های روزانه	کاهش اعتماد, مشکلات حقوقی	سیاست‌های حریم خصوصی, رمزنگاری
هماهنگی حملات	شبکه‌های ارتباطی	اختلال در سرویس, نقض امنیت	پایش ترافیک, نظارت مستمر
سیاه چاله	شبکه‌های ارتباطی	قطع ارتباط, اختلال در سرویس	تشخیص ناهنجاری, فیلتر ترافیک
حمله DDoS	شبکه‌های ارتباطی	قطع سرویس, اختلال عملیاتی	فایروال پیشرفته, محدودسازی ترافیک
باج افزارهای مبتنی بر هوش مصنوعی	شبکه‌های ارتباطی	قطع سرویس, زیان مالی	پشتیبان‌گیری منظم, فایروال پیشرفته

جدول ۸: حملات هوش مصنوعی تخاصمی در حوزه سلامت

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
استخراج مدل	مدل‌های هوش مصنوعی، APIها	سوءاستفاده از مدل	محدود کردن دسترسی به APIها، رصد ترافیک، بازنگری مدل‌ها، مسدود کردن دسترسی‌های غیرمجاز
بازسازی داده	داده‌های امنیتی و آمار فعالیت‌های روزانه	نقض حریم خصوصی, بازسازی غیرمجاز	پنهان‌سازی داده‌ها, رمزنگاری
استنتاج ویژگی	داده‌های امنیتی و آمار فعالیت‌های روزانه	نقض حریم خصوصی, سرقت اطلاعات	رمزنگاری داده‌ها, محدودسازی دسترسی
مسمومیت مدل هوش مصنوعی	فناوری‌های اختصاصی	کاهش دقت مدل, نقض امنیت	فیلتر داده ورودی, نظارت بر مدل
مسمومیت با برچسب تمیز	فناوری‌های اختصاصی	کاهش دقت مدل, نقض امنیت	اعتبارسنجی داده, نظارت بر مدل
مسمومیت درب پشتی	فناوری‌های اختصاصی	نقض امنیت, سرقت مالکیت فکری	کنترل دسترسی به کد, رمزنگاری
مسمومیت در دسترس پذیری	فناوری‌های اختصاصی	کاهش دسترسی, اختلال در سرویس	نظارت بر مدل, اعتبارسنجی داده

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
گریز یا نمونه های متخاصم	APIهای نظارتی و مدیریتی، فیلترهای امنیتی	فریب سامانه امنیتی، اختلال پیکربندی	آموزش مدل‌های هوش مصنوعی با داده‌های متنوع، محدود کردن دسترسی به APIها با کلیدهای امن، بررسی ورودی‌های API، رصد الگوهای غیرعادی ترافیک، به‌روزرسانی APIها، مسدود کردن درخواست‌های مشکوک
تزریق مستقیم و غیرمستقیم درخواست	فرم‌های نظرسنجی، بخش نظرات	تغییر محتوا، سرقت داده	پایش درخواست‌ها، رمزهای امنیتی، رصد ترافیک، مسدود کردن درخواست‌های غیرمجاز، بازنگری سامانه

۴-۴ شناسایی و مقابله با حملات مبتنی بر هوش مصنوعی در حوزه خبرگزاری‌ها و وبگاه‌های خبری

جدول ۹: حملات هوش مصنوعی تهاجمی در حوزه وبگاه‌های خبری

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تحلیل ترافیک شبکه	درخواست‌های HTTP/HTTPS	سرقت اطلاعات کاربر	HTTPS، دیوار آتش پیشرفته، رصد ترافیک، مسدود کردن ترافیک مشکوک، بازنگری تنظیمات شبکه
تشخیص آسیب‌پذیری	CMS، ابزارهای تعاملی	نفوذ به سامانه	به‌روزرسانی نرم‌افزارها، محدود کردن دسترسی، رصد ترافیک، رفع نقاط ضعف، بازنگری تنظیمات برنامه‌ها
پروفایل‌سازی هوشمند هدف‌محور	داده‌های کاربران، پروفایل‌ها	سوءاستفاده از اطلاعات شخصی	رمزنگاری داده‌ها، محدود کردن دسترسی، رصد ترافیک، بازنگری دسترسی‌ها، اصلاح داده‌های افشاشده
تحلیل رفتاری برای یافتن راه‌های جدید سوءاستفاده	داده‌های رفتاری کاربران	سوءاستفاده از الگوها	رمزنگاری داده‌های رفتاری، محدود کردن دسترسی، رصد رفتار کاربران، بازنگری دسترسی‌ها، اصلاح داده‌های افشاشده

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
رمزشکنی	پایگاه داده خبری، داده‌های مخاطبان	افشای اطلاعات حساس	رمزنگاری AES-256، محدود کردن دسترسی، رصد ترافیک، بازنگری الگوریتم‌های رمزنگاری، مسدود کردن دسترسی‌ها
مرد میانی	درخواست‌های HTTPS	سرقت یا دستکاری داده	HTTPS، رمزهای یک‌بارمصرف، بررسی گواهینامه‌های SSL/TLS، رصد ترافیک، بازنگری گواهینامه‌ها، مسدود کردن ترافیک مشکوک
تولید هوشمند نظرات جعلی	نظرسنجی‌ها، گزارش‌های مردمی	تحریف اطلاعات عمومی	تشخیص محتوای جعلی، CAPTCHAهای پیشرفته، احراز هویت کاربران، بررسی ترافیک، حذف نظرات جعلی، بازنگری سامانه‌های تعاملی
حمله به CAPTCHA	فرم‌های ثبت‌نام و ورود	دسترسی غیرمجاز	CAPTCHAهای پیشرفته، محدود کردن تلاش‌ها، رصد ترافیک، بازنگری CAPTCHA ها، مسدود کردن حساب‌های مشکوک
تولید خودکار اطلاعات نادرست	محتوای خبری، تصاویر، ویدیوها	فریب کاربران، تغییر افکار عمومی	ابزارهای تشخیص محتوای جعلی، آموزش کاربران، پایش محتوای ارسالی، بررسی رفتار کاربران، رصد ترافیک، حذف محتوای جعلی، بهبود الگوریتم‌های پایش
فیشینگ هدفمند	اطلاعات کاربران (ایمیل، پیام‌ها)	سرقت اطلاعات کاربر	آموزش کاربران، احراز هویت دو مرحله‌ای، پایش ایمیل‌ها، رصد ترافیک، مسدود کردن حساب‌های مهاجم، بازنگری امنیت حساب‌ها
تزریق اسکریپت‌های بین سایتی	فرم‌های ثبت‌نام، CMS	سرقت اطلاعات، تغییر محتوا	پایش ورودی‌ها، آموزش توسعه‌دهندگان، رصد ترافیک، حذف کدهای مخرب، بازنگری امنیت CMS
تزریق SQL	پایگاه داده خبری	سرقت یا تغییر اطلاعات	رمزنگاری پایگاه داده، رصد ترافیک، بازیابی از نسخه‌های پشتیبان، بازنگری امنیت پایگاه داده

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تهدیدهای پایدار پیشرفته	سرورها، گزارش‌های امنیتی، گواهینامه‌های SSL/TLS	قطعی سایت، سرقت داده، نقض اعتماد	رمزنگاری AES-256، احراز هویت دو مرحله‌ای، به‌روزرسانی نرم‌افزارها، جداسازی شبکه‌ها، تحلیل گزارش‌های امنیتی، سامانه تشخیص نفوذ، نسخه پشتیبان روزانه، قطع دسترسی‌های مشکوک
بدافزارهای خودآموز	سرورها، داده‌های امنیتی	تخریب سرور، سرقت اطلاعات	دیوار آتش وب، به‌روزرسانی نرم‌افزارها، احراز هویت دو مرحله‌ای، رصد ترافیک، بررسی رفتار کارکنان، حذف بدافزار، بازیابی سرورها
مهندسی معکوس	خدمات هوش مصنوعی (خلاصه‌سازی، ترجمه)	بازسازی مدل‌ها	محدود کردن دسترسی به APIها، رصد ترافیک، بازنگری مدل‌ها، مسدود کردن دسترسی‌های غیرمجاز
کانال جانبی	رفتار سرورها (زمان پاسخ، مصرف منابع)	کشف داده‌های امنیتی	رمزنگاری داده‌ها، احراز هویت قوی، رصد ترافیک، بررسی عملکرد سرور، بهینه‌سازی پاسخ‌گویی، بازنگری دسترسی‌ها
بدافزارهای مبتنی بر هوش مصنوعی برای فرار ازتشخیص	سامانه‌های تحویل محتوا (HTTP/HTTPS CDN)	اختلال یا سرقت محتوا	دیوار آتش وب، رمزنگاری ترافیک، رصد ترافیک، حذف بدافزار، بازنگری سامانه‌های تحویل محتوا
فرار از سامانه تشخیص نفوذ شبکه	دیوار آتش، گزارش‌های امنیتی	حذف داده، نقض امنیت	به‌روزرسانی دیوار آتش، دیوار آتش وب، بررسی ترافیک شبکه، رصد فعالیت کارکنان، بازسازی سامانه‌های امنیتی
سازگارسازی حمله	سرورها، APIهای نظارتی	قطعی خدمات، سرقت داده	احراز هویت قوی برای APIها، رمزنگاری ترافیک، سامانه تشخیص نفوذ پیشرفته، رصد ترافیک، مسدود کردن ترافیک مخرب، بازسازی سامانه
سرقت اعتبارنامه	نام کاربری، رمز عبور	سرقت حساب کاربری	رمزنگاری رمزها، رمزهای پیچیده، احراز هویت دو مرحله‌ای، رمزهای یک‌بارمصرف، رصد ترافیک، بازنشانی رمزها، مسدود کردن حساب‌های مهاجم

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
یافتن مسیر	جداول مسیریابی، تنظیمات DNS	هدایت ترافیک به مقصد مخرب	محدود کردن دسترسی به تنظیمات شبکه، رمزنگاری پیکربندی‌ها، رصد تغییرات DNS، بازنگری تنظیمات، بازیابی از نسخه‌های پشتیبان
حرکت جانبی	سرورها، روترها، دیوار آتش	دسترسی غیرمجاز	احراز هویت چندمرحله‌ای، جداسازی شبکه‌ها، رصد شبکه، مسدود کردن دسترسی‌های غیرمجاز، بازسازی سامانه
دستکاری رکوردها	پایگاه داده خبری	افزودن اخبار جعلی	رمزنگاری پایگاه داده، محدود کردن دسترسی، رصد تغییرات غیرمجاز، بازیابی از نسخه‌های پشتیبان، بازنگری دسترسی‌ها
حمله DDoS	سرورها، ابزارهای نظارت شبکه	قطعی خدمات	دیوار آتش وب، افزایش ظرفیت سرورها، رصد ترافیک، مسدود کردن ترافیک مخرب، استفاده از CDN
باج افزارهای مبتنی بر هوش مصنوعی	سرورها، گزارش‌های امنیتی	قفل داده، قطعی خدمات	رمزنگاری قوی، جداسازی شبکه‌ها، آموزش کارکنان، رصد ترافیک، ابزارهای ضدباج‌افزار، بازیابی از نسخه‌های پشتیبان، قطع دسترسی‌های مشکوک

جدول ۱۰: حملات هوش مصنوعی تخاصمی در حوزه وبگاه‌های خبری

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
استخراج مدل	مدل‌های هوش مصنوعی، APIها	سوءاستفاده از مدل	محدود کردن دسترسی به APIها، رصد ترافیک، بازنگری مدل‌ها، مسدود کردن دسترسی‌های غیرمجاز
بازسازی داده	پایگاه داده خبری، پروفایل‌های کاربران	سرقت اطلاعات حساس	رمزنگاری پایگاه داده، محدود کردن دسترسی، رصد ترافیک، بازیابی از نسخه‌های پشتیبان، بازنگری دسترسی‌ها
استنتاج عضویت	پروفایل‌ها، علایق کاربران	نقض حریم خصوصی	رمزنگاری AES-256، احراز هویت چندمرحله‌ای، رصد ترافیک، بازنگری دسترسی‌ها، به‌روزرسانی سیاست‌های امنیتی

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
استنتاج ویژگی	گزارش‌ها و تنظیمات امنیتی، گواهینامه‌های SSL/ TLS	افشای تنظیمات امنیتی	رمزنگاری گزارش‌ها، محدود کردن دسترسی، رصد گزارش‌ها، بررسی رفتار کارکنان، بازنگری دسترسی‌ها، اصلاح داده‌های افشاشده
مسمومیت مدل هوش مصنوعی	مدل‌های تحلیل گزارش امنیتی، تنظیمات سرور	تغییر رفتار امنیتی، نقض عملکرد	بررسی و پایش داده‌های ورودی، محدود کردن دسترسی، رصد تغییرات غیرعادی گزارش‌ها، بررسی رفتار کارکنان، بازیابی از نسخه‌های پشتیبان، بازنگری دسترسی‌ها
مسمومیت با برچسب تمیز	الگوریتم‌های پیشنهاد یا پایش محتوا	پیشنهاد محتوای جعلی	بررسی داده‌های ورودی، محدود کردن دسترسی به الگوریتم‌ها، رصد خروجی الگوریتم‌ها، بازآموزش الگوریتم‌ها، بازیابی از نسخه‌های پشتیبان
مسمومیت هدفمند	الگوریتم‌های پیشنهاد یا پایش محتوا	تغییر خروجی الگوریتم	بررسی داده‌های ورودی، محدود کردن دسترسی، رصد خروجی الگوریتم‌ها، بازآموزش الگوریتم‌ها، بازیابی از نسخه‌های پشتیبان
مسمومیت در دسترس پذیری	اشتراک‌های پولی، محتوای اختصاصی	مسدود شدن دسترسی کاربر	احراز هویت برای خدمات، به‌روزرسانی نرم‌افزارها، رصد ترافیک، بازیابی خدمات، افزایش ظرفیت سرورها
گریز یا نمونه های متخاصم	APIهای نظارتی و مدیریتی، فیلترهای امنیتی	فریب سامانه امنیتی، اختلال پیکربندی	آموزش مدل‌های هوش مصنوعی با داده‌های متنوع، محدود کردن دسترسی به APIها با کلیدهای امن، بررسی ورودی‌های API، رصد الگوهای غیرعادی ترافیک، به‌روزرسانی APIها، مسدود کردن درخواست‌های مشکوک
تزریق مستقیم و غیرمستقیم درخواست	فرم‌های نظرسنجی، بخش نظرات	تغییر محتوا، سرقت داده	پایش درخواست‌ها، رمزهای امنیتی، رصد ترافیک، مسدود کردن درخواست‌های غیرمجاز، بازنگری سامانه

۴-۵ شناسایی و مقابله با حملات مبتنی بر هوش مصنوعی در حوزه پیام‌رسان‌های همراه

جدول ۱۱: حملات هوش مصنوعی تهاجمی در حوزه پیام‌رسان‌های همراه

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تحلیل ترافیک شبکه	الگوهای ارتباطی، رفتارهای کاربران	مهاجمان الگوهای ترافیک پیام‌رسان، مانند زمان‌بندی یا حجم پیام‌ها، را تحلیل می‌کنند تا اطلاعات حساس در مورد رفتارها یا ارتباطات کاربران را استنباط کنند.	رمزنگاری متاداده، محدود کردن دسترسی. نظارت بر الگوهای ترافیک مشکوک. محدود کردن دسترسی، اطلاع‌رسانی.
تشخیص آسیب‌پذیری	منابع فناوری، داده امنیتی	مهاجمان پلتفرم پیام‌رسان را برای یافتن آسیب‌پذیری‌های شناخته‌شده در نرم‌افزار یا پیکربندی آن اسکن می‌کنند تا بتوانند به‌طور غیرمجاز دسترسی پیدا کنند، بدافزار نصب کنند یا خدمات را مختل کنند.	انجام ممیزی‌های امنیتی منظم، به‌روزرسانی نرم‌افزارها. استفاده از سیستم‌های تشخیص نفوذ برای شناسایی اسکن‌های غیرعادی. رفع سریع آسیب‌پذیری‌ها، محدود کردن دسترسی‌های غیرمجاز.
پروفایل‌سازی هوشمند هدف‌محور	اطلاعات رفتاری کاربران، تاریخچه پیام‌ها، موقعیت مکانی و علایق آن‌ها	مهاجمان با جمع‌آوری و تحلیل داده‌های کاربران، پروفایل دقیقی از آن‌ها ایجاد می‌کنند که می‌تواند برای حملات هدفمند مانند فیشینگ شخصی‌سازی‌شده یا مهندسی اجتماعی استفاده شود.	محدودسازی جمع‌آوری داده‌های غیرضروری، ناشناس‌سازی اطلاعات، کنترل‌های دقیق حریم خصوصی برای کاربران. نظارت بر فعالیت‌های مشکوک در سطح تحلیل رفتار. اطلاع‌رسانی به کاربران درباره دسترسی به داده‌های حساس.
تحلیل رفتاری برای یافتن راه‌های جدید سوءاستفاده	منابع فناوری، خدمات اصلی	مهاجمان از تکنیک‌های تحلیل رفتاری برای شناسایی الگوها یا ناهنجاری‌ها در رفتار کاربران پیام‌رسان استفاده می‌کنند تا از آن‌ها برای اهداف مخرب مانند هدف‌گیری کاربران برای فیشینگ یا سایر حملات بهره‌برداری کنند.	رمزنگاری داده‌ها، آموزش کاربران.

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
پیش‌بینی نتایج حملات	منابع فناوری، زیرساخت فیزیکی	مهاجمان از داده‌های پیام‌رسان، مانند الگوهای رفتاری کاربران یا محتوای پیام، برای پیش‌بینی احتمال یا نتیجه حملات آینده استفاده می‌کنند تا حملات خود را مؤثرتر برنامه‌ریزی یا اجرا کنند.	تصادفی‌سازی پاسخ‌ها، رمزنگاری داده‌ها.
رمزشکنی	داده‌های کاربر، دارایی‌های معنوی	مهاجمان از تکنیک‌های مختلف، مانند بهره‌برداری از آسیب‌پذیری‌های الگوریتم‌های رمزنگاری، به‌دست آوردن کلیدهای رمزنگاری یا رهگیری ارتباطات، برای رمزگشایی و خواندن محتوای پیام‌های رمزنگاری‌شده در پیام‌رسان استفاده می‌کنند.	داده‌های کاربر، دارایی‌های معنوی
مرد میانی	پروتکل‌های ارتباطی، داده‌های احراز هویت	مهاجمان ارتباطات پیام‌رسان را رهگیری می‌کنند و خود را بین فرستنده و گیرنده قرار می‌دهند تا پیام‌ها را استراق سمع کرده یا تغییر دهند.	استفاده از رمزنگاری قوی، سنجش گواهینامه‌ها.
تولید خودکار دامنه	داده‌های کاربر، دارایی‌های معنوی	مهاجمان از الگوریتم‌های تولید دامنه برای ایجاد تعداد زیادی دامنه استفاده می‌کنند که برای میزبانی محتوای مخرب یا خدماتی مانند لینک‌های فیشینگ یا دانلود بدافزار از طریق پیام‌رسان توزیع می‌شوند.	استفاده از خدمات شهرت دامنه، آموزش کاربران.
حمله به CAPTCHA	منابع فناوری، خدمات اصلی	مهاجمان از تکنیک‌هایی مانند حل‌کننده‌های خودکار، حل‌کننده‌های انسانی یا بهره‌برداری از آسیب‌پذیری‌های سیستم کپچا برای دور زدن چالش‌های کپچا در پیام‌رسان استفاده می‌کنند تا فعالیت‌های مخرب مانند اسپم، ایجاد حساب یا سایر حملات خودکار را انجام دهند.	منابع فناوری، خدمات اصلی

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تولید خودکار اطلاعات نادرست	داده‌ها و اطلاعات	مهاجمان از سیستم‌های خودکار مانند ربات‌ها یا هوش مصنوعی برای تولید و توزیع اطلاعات نادرست یا گمراه‌کننده از طریق پیام‌رسان استفاده می‌کنند تا کاربران را فریب دهند یا افکار عمومی را دستکاری کنند.	فیلتر کردن محتوای مشکوک، آموزش کاربران.
فیشینگ هدفمند و شخصی سازی‌شده	منابع فناوری، خدمات اصلی، داده‌های کاربر	مهاجمان پیام‌های شخصی‌سازی‌شده‌ای را از طریق پیام‌رسان ارسال می‌کنند که به نظر می‌رسد از منبعی معتبر هستند، تا کاربران را فریب دهند که اطلاعات حساس را فاش کنند یا روی لینک‌های مخرب کلیک کنند.	آموزش کاربران برای شناسایی پیام‌های مشکوک، استفاده از احراز هویت دو مرحله‌ای. فیلتر کردن پیام‌های مشکوک با استفاده از هوش مصنوعی، نظارت بر ترافیک غیرعادی. مسدود کردن حساب‌های متخلف، اطلاع‌رسانی به کاربران و بازیابی اطلاعات سرقت‌شده.
جعل	داده‌ها و اطلاعات، کاربران	مهاجمان با جعل اطلاعات فرستنده، مانند شماره تلفن یا نام کاربری، در پیام‌رسان، کاربران را فریب می‌دهند تا اطلاعات حساس را فاش کنند یا اقدامات غیرمجازی انجام دهند.	استفاده از امضای دیجیتال، رمزنگاری پیام‌ها.
سرقت هویت	خدمات امنیتی، داده‌های کاربر	مهاجمان هویت کاربران را در پیام‌رسان سرقت می‌کنند، اغلب با به‌دست آوردن اطلاعات ورود یا سایر اطلاعات شناسایی، تا به‌عنوان آن‌ها عمل کرده یا فعالیت‌های جعلی انجام دهند.	رمزنگاری داده‌ها، استفاده از احراز هویت دو مرحله‌ای.
تزریق اسکریپت‌های بین سایتی	داده‌ها و اطلاعات، رابطهای کاربری	مهاجمان اسکریپت‌های مخرب را در پیام‌ها یا محتوای پیام‌رسان تزریق می‌کنند که در مرورگر کاربران اجرا شده و امکان سرقت داده‌ها یا انجام اقدامات غیرمجاز را فراهم می‌کند.	اعتبارسنجی ورودی‌ها، استفاده از سیاست امنیتی محتوا.
تزریق SQL	داده‌ها و اطلاعات	مهاجمان کد SQL مخرب را در ورودی‌های کاربر پیام‌رسان تزریق می‌کنند که توسط پایگاه داده پلتفرم اجرا شده و امکان دسترسی، تغییر یا حذف داده‌ها را فراهم می‌کند.	استفاده از دستورات آماده ، اعتبارسنجی ورودی.

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
جعل عمیق	خدمات اصلی، داده‌های کاربری	مهاجمان از هوش مصنوعی برای تولید محتوای صوتی یا تصویری جعلی استفاده می‌کنند که از طریق پیام‌رسان توزیع می‌شود تا کاربران را فریب دهند یا اطلاعات نادرست منتشر کنند.	استفاده از سیستم‌های تشخیص جعلی، آموزش کاربران.
بازپخش	منابع فناوری، خدمات اصلی	مهاجمان پیام‌های قبلی را در پیام‌رسان ضبط و بازپخش می‌کنند تا سیستم یا کاربران را فریب دهند که معتقدند پیام‌ها جدید یا معتبر هستند، و به این ترتیب دسترسی غیرمجاز یا اقدامات دیگر را انجام می‌دهند.	استفاده از رمزنگاری قوی، اعتبارسنجی پیام‌ها.
تهدیدهای پایدار پیشرفته	منابع فناوری، دارایی‌های معنوی، داده‌ها و اطلاعات	مهاجمان از تکنیک‌های پیچیده برای دسترسی غیرمجاز به سیستم‌های پیام‌رسان و حفظ دسترسی طولانی‌مدت بدون شناسایی استفاده می‌کنند تا داده‌ها را سرقت کنند یا خدمات را مختل کنند.	تقسیم‌بندی شبکه، آموزش کارکنان.
توزیع بدافزار	خدمات اشتراک‌گذاری فایل، برنامه‌های کاربردی	مهاجمان لینک‌ها یا پیوست‌های مخرب را از طریق پیام‌رسان ارسال می‌کنند یا از آسیب‌پذیری‌های پلتفرم برای توزیع بدافزار به دستگاه‌های کاربران بهره‌برداری می‌کنند.	فیلتر کردن پیوست‌ها و لینک‌های مشکوک، آموزش کاربران.
کلون‌سازی	منابع فناوری، دارایی‌های معنوی	مهاجمان مدل‌های یادگیری ماشین پیام‌رسان را کپی می‌کنند تا از آن‌ها برای فعالیت‌های غیرمجاز یا مخرب، مانند دور زدن اقدامات امنیتی یا شبیه‌سازی رفتار پلتفرم، استفاده کنند.	مبهم‌سازی مدل، محدود کردن دسترسی.
مهندسی معکوس	الگوریتم‌های اختصاصی	مهاجمان کد، پروتکل‌ها یا باینری‌های پیام‌رسان را تحلیل می‌کنند تا عملکرد داخلی آن را درک کرده و آسیب‌پذیری‌ها را پیدا کنند، پلتفرم را شبیه‌سازی کنند یا سوءاستفاده‌هایی را توسعه دهند.	مبهم‌سازی کد، استفاده از تکنیک‌های ضد دیباگ.
کانال جانبی	داده‌های کاربر	مهاجمان از اطلاعات کانال جانبی، مانند الگوهای زمان‌بندی یا مصرف انرژی، از پیام‌رسان برای استنباط اطلاعات حساس در مورد پیام‌ها یا فعالیت‌های کاربران بهره‌برداری می‌کنند.	استفاده از الگوریتم‌های زمان ثابت، سپر الکترومغناطیسی.

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
بدافزارهای مبتنی بر هوش مصنوعی برای فرار از تشخیص	داده‌ها و اطلاعات، کاربران	مهاجمان جاسوس‌افزار را از طریق پیام‌رسان، اغلب از طریق لینک‌ها یا پیوست‌های مخرب، توزیع می‌کنند تا به دستگاه‌های کاربران دسترسی غیرمجاز پیدا کرده و فعالیت‌های آن‌ها، از جمله پیام‌ها، را نظارت کنند.	استفاده از آنتی‌ویروس، مدیریت دقیق دسترسی‌ها.
فرار از سامانه تشخیص نفوذ شبکه	منابع فناوری، خدمات امنیتی، داده‌های امنیتی	مهاجمان از تکنیک‌هایی برای دور زدن سیستم‌های تشخیص نفوذ مبتنی بر هوش مصنوعی پیام‌رسان استفاده می‌کنند، مانند دستکاری ورودی‌ها یا استفاده از نمونه‌های متخاصم، تا فعالیت‌های مخرب خود را بدون شناسایی انجام دهند.	به‌روزرسانی مدل، استفاده از تکنیک‌های دفاعی.
سازگارسازی حمله	منابع فناوری، خدمات اصلی	مهاجمان از تکنیک‌های تطبیقی مانند یادگیری ماشین یا هوش مصنوعی برای تغییر استراتژی‌های حمله خود بر اساس اقدامات امنیتی پیام‌رسان یا پاسخ‌های کاربران استفاده می‌کنند تا حملات خود را مؤثرتر کنند.	استفاده از لایه‌های امنیتی چندگانه، به‌روزرسانی منظم.
سرقت رمز عبور	خدمات امنیتی، داده‌های کاربر	مهاجمان از روش‌هایی مانند فیشینگ، بدافزار یا بهره‌برداری از آسیب‌پذیری‌های پلتفرم برای سرقت رمزهای عبور کاربران پیام‌رسان استفاده می‌کنند تا به حساب‌های آن‌ها دسترسی غیرمجاز پیدا کنند.	استفاده از احراز هویت دو مرحله‌ای، آموزش کاربران برای استفاده از رمزهای قوی.
استخراج داده‌ها	داده‌ها و اطلاعات	مهاجمان از تکنیک‌هایی مانند رهگیری ارتباطات، بهره‌برداری از آسیب‌پذیری‌های پلتفرم یا استفاده از بدافزار برای استخراج داده‌های حساس از پیام‌رسان یا دستگاه‌های کاربران استفاده می‌کنند.	رمزنگاری داده‌ها، محدود کردن دسترسی. نظارت بر انتقال داده‌های غیرعادی. مسدود کردن دسترسی‌های غیرمجاز، بازیابی داده‌ها.
هماهنگی حملات	منابع فناوری، خدمات اصلی	مهاجمان اقدامات خود را از طریق پیام‌رسان هماهنگ می‌کنند یا از پلتفرم به‌عنوان بخشی از یک حمله چندوجهی برای دستیابی به اهدافی مانند سرقت داده‌ها یا اختلال در خدمات استفاده می‌کنند.	محدود کردن نرخ درخواست‌ها، استفاده از احراز هویت قوی.

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
دستکاری رکوردها	داده‌ها و اطلاعات	مهاجمان به سوابق یا لاگ‌های پیام‌رسان دسترسی پیدا کرده و آن‌ها را تغییر می‌دهند تا محتوای پیام‌ها، فرستنده یا سایر جزئیات را برای اهداف مخرب مانند پنهان کردن ردپا یا انتشار اطلاعات نادرست دستکاری کنند.	استفاده از امضای دیجیتال، محدود کردن دسترسی.
حمله DDoS	پروتکل‌های شبکه، مدیریت ترافیک	مهاجمان با استفاده از دستگاه‌های متعدد درگیر، حجم زیادی از پیام‌ها یا درخواست‌ها را به سرورهای پیام‌رسان ارسال می‌کنند تا آن‌ها را بیش از حد بارگذاری کرده و خدمات را برای کاربران قانونی مختل کنند.	استفاده از خدمات کاهش DDoS، محدود کردن نرخ درخواست‌ها.

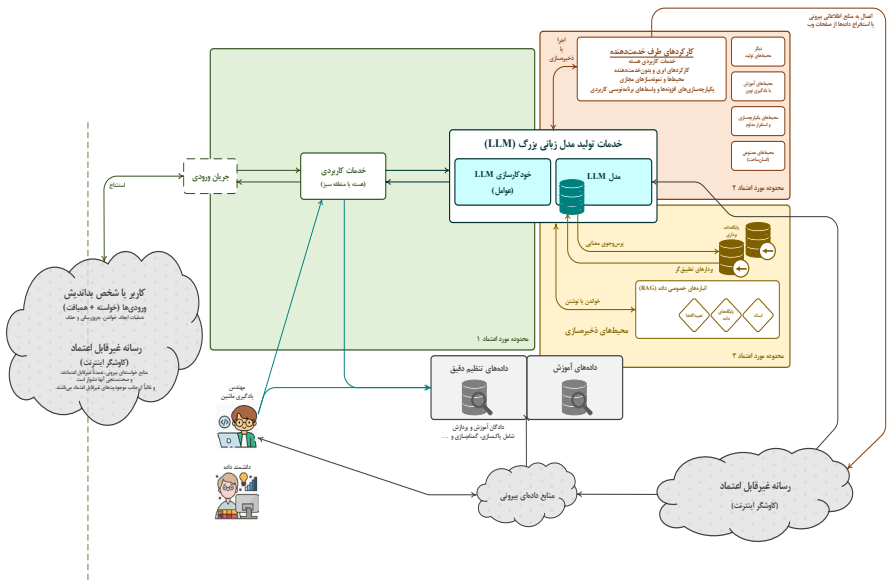
جدول ۱۲: حملات هوش مصنوعی تخاصمی در حوزه پیام‌رسان‌های همراه

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
استخراج مدل	منابع فناوری، دارایی‌های معنوی	مهاجمان مدل‌های یادگیری ماشین پیام‌رسان را استخراج می‌کنند تا اطلاعات حساس در مورد ساختار یا داده‌های آموزشی آن‌ها را به‌دست آورند، که می‌تواند برای توسعه حملات یا سوءاستفاده‌های بیشتر استفاده شود.	مبهم‌سازی مدل، استفاده از حریم خصوصی تفاضلی.
بازسازی داده	منابع فناوری، خدمات اصلی	مهاجمان از تکنیک‌هایی برای بازسازی یا بازیابی داده‌هایی که کاربران معتقدند حذف شده یا امن هستند، مانند دسترسی به نسخه‌های پشتیبان، لاگ‌ها یا سایر منابع داده در پیام‌رسان استفاده می‌کنند.	رمزنگاری داده‌ها، محدود کردن دسترسی.
استنتاج ویژگی	منابع فناوری، داده‌های کاربر	مهاجمان از داده‌های پیام‌رسان، مانند محتوای پیام، فرکانس یا الگوها، برای استنباط اطلاعات حساس در مورد کاربران، مانند مکان، علایق یا سایر جزئیات شخصی استفاده می‌کنند.	استفاده از حریم خصوصی تفاضلی، رمزنگاری متاداده.
مسمومیت مدل هوش مصنوعی	منابع فناوری، داده‌های کاربر	مهاجمان مدل‌های هوش مصنوعی پیام‌رسان را مسموم می‌کنند، یا با خراب کردن داده‌های آموزشی یا دستکاری مستقیم مدل، تا باعث شوند اشتباهاتی انجام دهند یا به‌صورت ناخواسته رفتار کنند، مانند طبقه‌بندی نادرست پیام‌ها یا کاربران.	اعتبارسنجی داده‌های آموزشی، نظارت مداوم.

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
مسمومیت درب پشتی	منابع فناوری، خدمات امنیتی، داده‌های کاربر	مهاجمان درب‌های پشتی را در داده‌ها یا سیستم‌های پیام‌رسان تزریق می‌کنند تا دسترسی غیرمجاز مداوم ایجاد کنند یا رفتار سیستم را به‌صورت مخفیانه دستکاری کنند.	اعتبارسنجی داده‌ها، استفاده از حریم خصوصی تفاضلی.
مسمومیت در دسترس پذیری	داده‌های کاربر، دارایی‌های معنوی	مهاجمان داده‌ها یا مکانیزم‌های کنترل دسترسی پیام‌رسان را مسموم می‌کنند، مانند تزریق ورودی‌های مخرب به پایگاه‌های داده یا بهره‌برداری از آسیب‌پذیری‌های سیستم احراز هویت، برای دسترسی غیرمجاز به حساب‌ها یا پیام‌های کاربران.	داده‌های کاربر، دارایی‌های معنوی
گریز یا نمونه های متخاصم	منابع فناوری، داده‌های کاربر	مهاجمان نمونه‌های متخاصم را به داده‌های پیام‌رسان اضافه می‌کنند که برای فریب مدل‌های یادگیری ماشین پلتفرم یا سیستم‌های خودکار طراحی شده‌اند تا اشتباهاتی مانند طبقه‌بندی نادرست پیام‌ها یا کاربران را ایجاد کنند و فعالیت‌های مخرب را ممکن سازند.	منابع فناوری، داده‌های کاربر
حمله مبتنی بر RAG	منابع فناوری، خدمات مبتنی بر هوش مصنوعی، داده‌های کاربر	مهاجمان از تکنیک‌های مبتنی بر بازیابی و تولید (RAG) برای دستکاری پاسخ‌های سیستم‌های هوش مصنوعی پیام‌رسان استفاده می‌کنند تا اطلاعات نادرست تولید کنند یا رفتار سیستم را تغییر دهند.	فیلتر کردن ورودی‌ها، نظارت مداوم.
تزریق مستقیم درخواست	منابع فناوری، داده‌های کاربر	مهاجمان درخواست‌های مخرب را مستقیماً به API یا سایر رابط‌های پیام‌رسان تزریق می‌کنند تا از آسیب‌پذیری‌ها بهره‌برداری کرده و دسترسی غیرمجاز، دستکاری داده‌ها یا سایر اقدامات مخرب را انجام دهند.	اعتبارسنجی ورودی، استفاده از دیوار آتش برنامه.
تزریق غیرمستقیم درخواست	منابع فناوری، داده‌های کاربر	مهاجمان درخواست‌های مخرب را از طریق روش‌های غیرمستقیم، مانند خدمات شخص ثالث درگیر یا بهره‌برداری از وابستگی‌های پلتفرم، به پیام‌رسان تزریق می‌کنند تا به اهداف مشابه تزریق مستقیم دست یابند.	اعتبارسنجی ورودی، استفاده از دیوار آتش برنامه. نظارت بر درخواست‌های مشکوک. رفع آسیب‌پذیری، اطلاع‌رسانی.

۴-۶ شناسایی و مقابله با حملات مبتنی بر هوش مصنوعی در حوزه گفتگوگرهای مبتنی بر مدل‌های زبانی بزرگ

در خصوص مدل‌های زبانی بزرگ، شناسایی و مقابله با حملات مبتنی بر هوش مصنوعی بر مبنای معماری مشخص شده در شکل ۱ صورت گرفته است.



شکل ۱: معماری کاربرد مدل‌های زبانی بزرگ

در ادامه بر مبنای معماری شکل ۱ حملات تهاجمی و تخصامی در حوزه مدل‌های زبانی بزرگ مشخص شده اند.

جدول ۱۳: حملات هوش مصنوعی تهاجمی در حوزه گفتگوگرهای مدل‌های زبانی بزرگ

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تحلیل ترافیک شبکه	جریان ورودی، کارکردهای طرف خدمت‌دهنده، محیطهای ذخیره‌سازی و داده‌ها	دسترسی و افشای اطلاعات حساس	پوشاندن ترافیک و افزودن نویز به حجم داده‌ها، استفاده از پروتکل‌های امن
تشخیص آسیب‌پذیری	خدمات کاربردی، کارکردهای طرف خدمت‌دهنده، خدمات تولید مدل زبانی بزرگ، محیطهای ذخیره‌سازی و داده‌ها	شناسایی نقص‌های امنیتی و اجرای موثرتر حملات	نظارت بر رفتارهای مشکوک در لاگ سامانه و پاسخ‌ها، استفاده از سیستم‌های تشخیص الگوهای اسکن

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
جاسوسی	محیط‌های ذخیره‌سازی و داده‌ها، خدمات کاربردی، منابع داده بیرونی و رسانه غیرقابل اعتماد، کارکردهای طرف خدمت‌دهنده	نظارت و دسترسی غیرمجاز اطلاعات	رمزنگاری تمام ارتباطات، کنترل دقیق قابلیت حافظه در مدل
پروفایل‌سازی هوشمند هدف‌محور	جریان ورودی، خدمات کاربردی، منابع داده بیرونی و رسانه غیرقابل اعتماد	طراحی حملات دقیق علیه تعاملات کاربر	بررسی ساختار درخواست برای شناسایی رفتار پروفایل‌سازی، محدودیت دسترسی به داده‌های حساس در خروجی
تحلیل رفتاری برای یافتن راه‌های جدید سوءاستفاده	محیط‌های ذخیره‌سازی و داده‌ها، کارکردهای طرف خدمت‌دهنده، خدمات کاربردی، جریان ورودی	تحلیل الگوهای رفتاری و استخراج نقاط آسیب‌پذیری برای افزایش اثربخشی بهتر حمله	بررسی درخواست‌هایی که ممکن است با هدف کشف قوانین رفتاری ارسال شده باشند، محدودسازی قابلیت‌های مدل در افشای اطلاعات ساختاری
پیش‌بینی نتایج حملات	کارکردهای طرف خدمت‌دهنده، خدمات کاربردی	بررسی رفتارها و بهبود سناریو حمله برای اثربخشی بیشتر	تولید خروجی‌های مدل به صورت شبه‌تصادفی و دارای تنوع، ارزیابی مداوم مدل برای کشف مسیرهای حمله پیش‌بینی‌شده
رمزشکنی	محیط‌های ذخیره‌سازی و داده‌ها، کارکردهای طرف خدمت‌دهنده، خدمات تولید مدل زبانی بزرگ	دسترسی غیرمجاز به داده‌های حساس و رمزنگاری شده	استفاده از الگوریتم‌های رمزنگاری استاندارد و به‌روز، رمزگذاری داده‌های حساس
کش‌کاوی	کارکردهای طرف خدمت‌دهنده، محیط‌های ذخیره‌سازی و داده‌ها	دسترسی غیرمجاز به اطلاعات کش	نظارت رفتار مدل برای تشخیص الگوهای دسترسی غیرعادی، استفاده از ایزوله‌سازی کش برای هر کاربر
مرد میانی	جریان ورودی، خدمات کاربردی، محیط‌های ذخیره‌سازی و داده‌ها	نشت اطلاعات، پیام ها و دستورات در حال انتقال	به‌کارگیری پروتکل‌های رمزنگاری قوی، کنترل نشست‌های کاربر
تولید خودکار دامنه	منابع داده بیرونی و رسانه غیرقابل اعتماد، کارکردهای طرف خدمت‌دهنده	ارتباط با سامانه‌های آلوده در سمت سرور و آلودگی مدل	مسدودسازی دامنه‌هایی با الگوهای مشکوک، استفاده از سامانه‌های مبتنی بر هوش مصنوعی برای شناسایی فعالیت‌های DNS غیرعادی
مبهم‌سازی بدافزار	کارکردهای طرف خدمت‌دهنده، خدمات تولید مدل زبانی بزرگ، محیط‌های ذخیره‌سازی و داده‌ها	فرار از سامانه‌های تشخیص و اعمال تغییرات ناخواسته در عملکرد مدل	تشخیص کدهای مخرب با سامانه‌های مبتنی بر هوش‌مصنوعی، تحلیل معنایی کدهای ارسال‌شده

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
تولید داده‌های مصنوعی	منابع داده و رسانه غیرقابل اعتماد، خدمات تولید مدل زبانی بزرگ، محیطهای ذخیره‌سازی و داده‌ها	انحراف رفتار مدل و سوگیری نتایج	شناسایی الگوهای غیرعادی در داده‌های آموزشی/ورودی، اعتبارسنجی داده‌ها پیش از ورود به آموزش مدل
تولید هوشمند نظرات جعلی	خدمات تولید مدل زبانی بزرگ، منابع داده بیرونی و رسانه غیرقابل اعتماد، جریان ورودی	انحراف و ایجاد توهم برای مدل	تحلیل سبک نوشتار به کمک هوش مصنوعی برای تشخیص تولید ماشینی، آموزش مدل تشخیص محتوای مصنوعی در کنار مدل اصلی
حمله به CAPTCHA	خدمات کاربردی	اعمال حملات خودکار	اعمال نرخ سخت‌گیرانه بر تلاش‌های CAPTCHA، استفاده از CAPTCHA مبتنی بر تعامل معنایی
تولید خودکار اطلاعات نادرست	خدمات تولید مدل زبانی بزرگ، منابع داده بیرونی و رسانه غیرقابل اعتماد	اختلال در روند آموزش و کاهش کیفیت عملکرد مدل	اعتبارسنجی محتوا به کمک منابع تاییدشده اطلاعاتی، آموزش مدل برای تشخیص نشانه‌های اطلاعات گمراه‌کننده
سیبل	جریان ورودی، منابع داده بیرونی و رسانه غیرقابل اعتماد، خدمات کاربردی	خطای تصمیم‌گیری سامانه بر اساس ورودی‌های جعلی	اعتبارسنجی هویت کاربران با سازوکارهایی مانند CAPTCHA، محدود کردن دسترسی
فیشینگ هدفمند	جریان ورودی، خدمات کاربردی، منابع داده بیرونی و رسانه غیرقابل اعتماد	سرقت اطلاعات حساس کاربر، نفوذ به سرویس‌ها	احراز هویت چند مرحله‌ای، تشخیص فیشینگ به کمک مدل زبانی ثانویه، به‌کارگیری مدل زبانی ثانویه برای تشخیص الگوهای غیرعادی، پیاده‌سازی DKIM برای تایید اصالت فرستنده
حدس رمز عبور	جریان ورودی، محیطهای ذخیره سازی و داده‌ها، خدمات کاربردی	دسترسی به اطلاعات حساس کاربران	آموزش کاربران برای اجتناب از اطلاعات شخصی در رمز عبور، احراز هویت چندمرحله‌ای
سرقت رمز عبور	منابع داده بیرونی و رسانه غیرقابل اعتماد، جریان ورودی	دسترسی به اطلاعات کاربران	احراز هویت چندمرحله‌ای، نظارت لاگین‌ها از دستگاه‌ها و مکان‌های غیرعادی
بروت فورس رمز عبور	خدمات کاربردی، جریان ورودی، محیطهای ذخیره‌سازی و داده‌ها	دسترسی به اطلاعات حساس کاربران	احراز هویت چندمرحله‌ای، رمزنگاری قوی رمز عبورها
جعل	جریان ورودی، خدمات کاربردی، منابع داده بیرونی و رسانه غیرقابل اعتماد	جعل هویت و دسترسی غیرمجاز	تحلیل سبک زبانی درخواست‌ها به کمک مدل زبانی، احراز هویت رفتاری

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
جعل بیومتریک	جریان ورودی، خدمات کاربردی	جعل هویت و دسترسی غیرمجاز	استفاده از بیومتریک مقاوم مانند تشخیص زنده بودن، بررسی ناهمخوانی سیگنال‌های بیومتریک با رفتار معمول
سرقت هویت	جریان ورودی، خدمات کاربردی	جعل هویت	احراز هویت چندمرحله‌ای، دریابی تغییرات ناگهانی در سبک تعامل کاربر
تزریق اسکریپت‌های بین سایتی	خدمات کاربردی، جریان ورودی، کارکردهای طرف خدمت‌دهنده	اجرای کد مخرب در سامانه	استفاده از CSP برای جلوگیری از اجرای اسکریپت‌های تزریقی، استفاده از مدل زبانی مجزا برای تشخیص الگوهای XSS
تزریق SQL	جریان ورودی، محیطهای ذخیره‌سازی و داده‌ها	دسترسی و تغییر داده‌های حساس	پیاده‌سازی WAF با تحلیل رفتاری،استفده از دستورات امن و پارامتری‌سازی شده
جعل عمیق	خدمات تولید مدل زبانی بزرگ، جریان ورودی، منابع داده بیرونی و رسانه غیرقابل اعتماد	انحراف مدل	واترمارک یا اثرانگشت دیجیتال، بررسی تطابق میان صوت و تصویر
بازپخش	جریان ورودی، خدمات کاربردی	اختلال واسط کاربری در اثر جعل درخواست‌های کاربران	احراز هویت چند مرحله‌ای، استفاده از توکن‌های یکبار مصرف
تهدیدهای پایدار پیشرفته	کارکردهای طرف خدمت‌دهنده، خدمات کاربردی، محیطهای ذخیره‌سازی و داده‌ها، خدمات تولید مدل زبانی بزرگ	نفوذ عمیق و ماندگار در سامانه، سرقت یا انحراف رفتار مدل	تفکیک محیطهای اجرا و حافظه مدل، تشخیص ناهنجاری‌های ظریف در سامانه به کمک سامانه‌های مبتنی بر هوش مصنوعی
توزیع بدافزار	خدمات کاربردی، کارکردهای طرف خدمت‌دهنده، منابع داده بیرونی و رسانه غیرقابل اعتماد	گسترش عملکرد بدافزار در سامانه	تحلیل هوشمند فایل‌ها و کدها، استقرار WAF برای تشخیص درخواست‌های حاوی بدافزار
بدافزارهای خودآموز	کارکردهای طرف خدمت‌دهنده، خدمات تولید مدل زبانی بزرگ، محیطهای ذخیره‌سازی و داده‌ها	فرار از سامانه‌های امنیتی از طریق یادگیری رفتارهای سامانه، تغییر رفتار مدل	شبیه‌ساز تهدید برای تست دفاع در برابر بدافزارها، نظارت کدهای در حال تحول با تحلیل رفتار در زمان اجرا
کلون‌سازی	جریان ورودی، منابع داده بیرونی و رسانه غیرقابل اعتماد، خدمات کاربردی	تصمیم‌گیری نادرست سامانه بر اساس اطلاعات جعلی وارد شده	احراز هویت چندمرحله‌ای، نگهداری لاگ‌های دقیق از فعالیت سامانه

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
مهندسی معکوس	خدمات تولید مدل زبانی بزرگ، کارکردهای طرف خدمت‌دهنده، محیط‌های ذخیره‌سازی و داده‌ها، خدمات کاربردی	نشت الگوریتم‌ها و مدل‌ها	نظارت بر الگوی درخواست‌های مشکوک، تنوع در پاسخ‌دهی برای جلوگیری از نتیجه‌گیری دقیق
کانال جانبی	خدمات تولید مدل زبانی بزرگ، محیط‌های ذخیره‌سازی و داده‌ها	دسترسی غیرمجاز به داده‌های حساس	اجرای مدل در محیط محافظت‌شده، نظارت دقیق تعاملات با منابع سیستمی
بدافزارهای مبتنی بر هوش مصنوعی برای فرار از تشخیص	محیط‌های ذخیره‌سازی و داده‌ها، کارکردهای طرف خدمت‌دهنده، خدمات کاربردی	فرار از سامانه‌های تشخیص و آسیب به عملکرد سامانه	اجرای کدهای تولیدشده در محیط ایزوله و نظارت‌شده، تحلیل الگوهای رفتاری مشکوک به کمک مدل‌های یادگیری ماشین
دور زدن سامانه‌های تشخیص تهدیدات داخلی	کارکردهای طرف خدمت‌دهنده، محیط‌های ذخیره‌سازی و داده‌ها، خدمات تولید مدل زبانی بزرگ	سرقت یا دستکاری اطلاعات، اختلال در فرآیند یادگیری مدل	تعریف baseline رفتاری دقیق و پویا برای هر نقش، شناسایی تغییرات نامحسوس در سبک درخواست‌ها
دور زدن فیلترهای ایمیل	جریان ورودی، منابع داده بیرونی و رسانه غیرقابل اعتماد	انتقال داده‌های مخرب	استفاده از مدل زبانی ثانویه برای تحلیل ایمیل‌های ورودی، مسدودسازی ایمیل‌های احتمالا مخرب براساس الگوهای پیش‌بینی‌شده
فرار از سامانه تشخیص نفوذ شبکه	کارکردهای طرف خدمت‌دهنده، محیط‌های ذخیره‌سازی و داده‌ها	دور زدن سامانه‌های امنیتی و نفوذ در سامانه	به‌کارگیری آموزش تخاصمی، استفاده از مدل زبانی مجزا برای تشخیص رفتارهای غیرعادی
سازگارسازی حمله	کارکردهای طرف خدمت‌دهنده، جریان ورودی، خدمات کاربردی، خدمات تولید مدل زبانی بزرگ	تحلیل الگوهای رفتاری و دفاعی و عبور از کنترل‌های امنیتی سامانه	تحلیل الگوی توالی درخواست‌ها برای شناسایی حملات تدریجی، آموزش متقابل سامانه به کمک شبیه‌سازی سناریو حمله
ثبت کلید ضمنی	جریان ورودی، محیط‌های ذخیره‌سازی و داده‌ها، کارکردهای طرف خدمت‌دهنده	تحلیل رفتار داخلی سامانه و دسترسی غیرمجاز به کلیدهای رمزنگاری شده	نظارت بر خروجی مدل، استفاده از مدل زبانی محافظ برای تحلیل درخواست
سرقت اعتبارنامه	جریان ورودی، محیط‌های ذخیره‌سازی و داده‌ها، کارکردهای طرف خدمت‌دهنده	دسترسی غیرمجاز به رمزها و توکن‌ها و ورود به سامانه‌های مدیریتی	احراز هویت چندمرحله‌ای، استفاده از فیلترهای ورودی برای شناسایی محتوای شبه‌اعتبارنامه

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
یافتن مسیر	کارکردهای طرف خدمت‌دهنده، خدمات کاربردی، جریان ورودی، محیطهای ذخیره‌سازی و داده‌ها	دسترسی به آسیب‌پذیری‌های سامانه و استخراج اطلاعات حساس	تحلیل توالی درخواست‌ها برای کشف تلاش نقشه‌برداری مسیر، پیاده‌سازی محدودیت نرخ در APIهای مربوط به مدل
حرکت جانبی	کارکردهای طرف خدمت‌دهنده، جریان ورودی، خدمات کاربردی	نفوذ مهاجم از یک سرویس به یک سرویس دیگر و گسترش حمله	تحلیل ارتباط سرویس‌ها، استقرار راه‌حل‌هایی مانند EDR
استخراج داده‌ها	کارکردهای طرف خدمت‌دهنده، محیطهای ذخیره‌سازی و داده‌ها	استخراج داده‌های حساس	پیاده‌سازی DLP مبتنی بر مدل زبانی بزرگ برای تشخیص داده‌های حساس، تعریف سیاست‌ها برای جلوگیری از افشای اطلاعات حساس
تغییر فیلم‌های نظارتی	منابع داده بیرونی و رسانه غیرقابل اعتماد، محیطهای ذخیره‌سازی و داده‌ها، جریان ورودی	گمراهی مدل و سامانه	تحلیل داده تصویری به کمک سامانه مبتنی بر هوش مصنوعی
حمله حریم خصوصی	محیطهای ذخیره‌سازی و داده‌ها	دسترسی غیرمجاز به اطلاعات حساس	فیلتر خروجی برای جلوگیری از انتشار اطلاعات هویتی، نظارت و تشخیص درخواست‌هایی که دارای نشانه اطلاعات کاوی هستند
هماهنگی حملات	همه دارایی‌ها	اعمال حملات هماهنگ به نقاط مختلف سامانه	ایزوله‌سازی مدل از سایر سرویس‌های آسیب‌پذیر، تحلیل زمان‌بندی و ترتیب رویدادها برای کشف الگوهای هماهنگ
سیاه چاله	جریان ورودی، کارکردهای طرف خدمت‌دهنده	اختلال در شبکه	استفاده از timeoutهای سخت‌گیرانه، تحلیل رفتاری به کمک مدل زبانی ثانویه
دستکاری رکوردها	محیطهای ذخیره‌سازی و داده‌ها، خدمات تولید مدل زبانی بزرگ	انحراف مدل در صورت تغییر رکوردهای داده‌های آموزشی	تحلیل الگوهای تغییر داده، بررسی کیفیت و اصالت داده
حمله DDoS	خدمات کاربردی، جریان ورودی	از کار افتادن سرویس‌ها	بهره‌گیری از CDN مقاوم به DDos، استفاده از سیستم‌های تشخیص رفتار بات‌نت با هوش‌مصنوعی
باج افزارهای مبتنی بر هوش مصنوعی	کارکردهای طرف خدمت‌دهنده، خدمات کاربردی، جریان ورودی	دسترسی به اطلاعات حساس	نظارت مداوم بر فعالیت‌های غیرعادی، تحلیل رفتاری به کمک هوش مصنوعی

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
دستکاری افکار عمومی	خدمات تولید مدل زبانی بزرگ، منابع داده بیرونی و رسانه غیرقابل اعتماد، جریان ورودی	تولید محتوای تحریف‌شده توسط مدل	تحلیل سبک نگارش و شباهت‌های معنایی برای کشف متون تولیدشده انبوه، استفاده از مدل زبانی مجزا برای تایید صحت محتوا
جمینگ	کارکردهای طرف خدمت‌دهنده، خدمات کاربردی، جریان ورودی	اختلال در سرویس‌دهی	بهره‌گیری از یادگیری ماشین تخاصمی برای شبیه‌سازی حمله، شناسایی ناسازگاری‌ها به کمک سامانه‌های مبتنی بر هوش مصنوعی

جدول ۱۴: حملات هوش مصنوعی تخاصمی در حوزه گفتگوگرهای مدل‌های زبانی بزرگ

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
استخراج مدل	خدمات تولید مدل زبانی بزرگ، کارکردهای طرف خدمت‌دهنده	کپی برداری و استنتاج ساختار مدل	اعتبارسنجی به جهت جلوگیری از درخواست‌های خصمانه، استفاده از تکنیک‌های جلوگیری از بیش‌برازش مدل
بازسازی داده	خدمات تولید مدل زبانی بزرگ، محیطهای ذخیره‌سازی و داده‌ها، کارکردهای طرف خدمت‌دهنده	بازسازی داده‌های حساس	رمزگذاری داده حساس، شناسایی الگوهای غیرمعمول در دسترسی به داده و استفاده از مدل
استنتاج عضویت	محیطهای ذخیره‌سازی و داده‌ها، خدمات تولید مدل زبانی بزرگ	افشای داده‌های حساس	کاهش بیش‌برازش با به‌کارگیری مجموعه داده متنوع، مبهم‌سازی اطلاعات قابل شناسایی برای حفاظت از اطلاعات حساس
استنتاج ویژگی	محیطهای ذخیره‌سازی و داده‌ها، خدمات تولید مدل زبانی بزرگ	افشای اطلاعات آماری و ویژگی‌های حساس مدل	کنترل سطح دسترسی به توابع تحلیلی یا آمارگونه در مدل، آموزش و به‌کارگیری یک مدل ساده‌تر با رفتار نزدیک به مدل اصلی
مسمومیت مدل هوش مصنوعی	خدمات تولید مدل زبانی بزرگ، محیطهای ذخیره‌سازی و داده‌ها، منابع داده بیرونی و رسانه غیرقابل اعتماد	مسمومیت و آلودگی مدل	اعتبارسنجی داده و مقایسه خروجی مدل با منابع قابل اعتماد، آموزش متخاصم مدل برای حساسیت کمتر به داده‌های مسموم
مسمومیت با برچسب تمیز	خدمات تولید مدل زبانی بزرگ، محیطهای ذخیره‌سازی و داده‌ها، منابع داده بیرونی و رسانه غیرقابل اعتماد	مسمومیت و آلودگی مدل	آموزش متخاصم برای مقاوم‌سازی عملکرد مدل، بهره‌گیری از بازخورد انسان برای شناسایی رفتار مخرب مدل
مسمومیت هدفمند	خدمات تولید مدل زبانی بزرگ، محیطهای ذخیره‌سازی و داده‌ها، منابع داده بیرونی و رسانه غیرقابل اعتماد	مسمومیت و آلودگی مدل	به‌کارگیری داده از منابع متنوع، طراحی تست مفهومی برای کشف انحرافات مدل

حمله	دارایی‌های هدف	تأثیر	راهکارهای مقابله
مسمومیت درب پشتی	خدمات تولید مدل زبانی بزرگ، محیطهای ذخیره‌سازی و داده‌ها، منابع داده بیرونی و رسانه غیرقابل اعتماد	مسمومیت و آلودگی مدل	اعتبارسنجی داده‌های آموزشی از منابع معتبر، آموزش مدل به کمک روش‌های خصمانه برای افزایش استحکام
مسمومیت در دسترس پذیری	خدمات تولید مدل زبانی بزرگ، محیطهای ذخیره‌سازی و داده‌ها، منابع داده بیرونی و رسانه غیرقابل اعتماد، خدمات کاربردی	مسمومیت و آلودگی مدل، کاهش دسترس‌پذیری سامانه	آموزش با مجموعه داده تمیز از منابع تاییدشده، افزایش استحکام مدل در برابر ورودی تحریف‌شده از طریق آموزش همراه با نمونه خصمانه
گریز یا نمونه های متخاصم	جریان ورودی، خدمات تولید مدل زبانی بزرگ، خدمات کاربردی	فریب و گمراهی مدل	استفاده از تنظیم دقیق مستحکم با داده‌های متخاصم، استفاده از آشکارساز مخصوص حملات متخاصم
حمله مبتنی بر RAG	محیطهای ذخیره‌سازی و داده‌ها، خدمات تولید مدل زبانی بزرگ، منابع داده بیرونی و رسانه غیرقابل اعتماد	تولید محتوای دستکاری‌شده و مخرب	اعتبارسنجی منابع بازیابی‌شده با سامانه‌های مبتنی بر هوش مصنوعی، به‌روزرسانی مدل به طور تکراری با کمک یادگیری تدریجی
تزریق مستقیم درخواست	جریان ورودی، خدمات تولید مدل زبانی بزرگ، خدمات کاربردی	تولید محتوای مخرب با ورود خواسته‌های دستکاری‌شده	اعتبارسنجی درخواست‌های ارسال شده به مدل، محدودکردن دسترسی کاربران بر اساس نقش
تزریق غیرمستقیم درخواست	محیطهای ذخیره‌سازی و داده‌ها، خدمات تولید مدل زبانی بزرگ، خدمات کاربردی	تولید محتوای مخرب با ورود خواسته‌های دستکاری‌شده	محدودسازی محتوای خارجی و کنترل منابع، اعتبارسنجی داده‌های برگرفته از منابع خارجی