Google

रिस्पॉन्सिबल मशीन लर्निंग के लिए दिशा-निर्देश

दिशा-निर्देश

एमएल मॉडल के आउटपुट के आधार पर कंट्रोल फ़्लो को बदलते समय सावधानी बरतें.

किसी मशीन लर्निंग प्रिमिटिव के आउटपुट को कंज़्यूम करने वाले वर्कफ़्लो को ऐसे हमले पैदा करने के लिए मॉडिफ़ाई किया जा सकता है जिनसे सुरक्षा करना मुश्किल है. मौजूदा सलाह के हिसाब से, एग्ज़ीक्यूशन वर्कफ़्लो में डेटा को इंजेस्ट नहीं किया जाना चाहिए और साफ़ तौर पर इसकी अपेक्षा नहीं की जाती है.

मॉडल और मॉडल मेटाडेटा को सीधे ऐक्सेस करने से बचें.

मशीन लर्निंग प्रिमिटिव के साथ सीधे तौर पर इंटरैक्ट करने से हमले की कई अलग-अलग सतहें बन सकती हैं; घुसपैठ, इन्वर्ज़न वगैरह. इन ऑब्जेक्ट को उपयोगकर्ताओं से दूर रखें और कंट्रोल वाले, केवल-अनुमत इंटरैक्शन की अनुमति दें. साथ ही, पक्का करें कि आउटपुट एक स्थापित सुविधा से जुड़ा हुआ है - जैसे डेटाबेस. उपयोगकर्ता को सीधे आउटपुट उपलब्ध न कराएं.

पक्का करें कि मॉडल, ट्रेनिंग और इंप्लीमेंटेशन के बाद दोनों पर इंटेग्रिटी चेक का इस्तेमाल किया जाए.

सॉफ़्टवेयर को बनाए और डिप्लॉय किए जाने के कई चरणों के दौरान, उसमें बदलाव किए जा सकते हैं, यही बात मशीन लर्निंग पर भी लागू होती है. जांचें और फिर से जांचें. केवल वही रन करें जिसके बारे में आपको पता हो कि वह अपेक्षित है: ऑपरेट करने से पहले प्रमाणित करें.

सुविधाओं की जटिलता से कंट्रोल की जटिलता पैदा हो सकती है.

इंटरैक्शन को छोटा, सटीक और काम तक सीमित रखें. बड़े, जटिल वर्कफ़्लो से सफल हमलों का जोखिम बहुत बढ़ जाता है.

ट्रेनिंग के लिए एक्सटेंशनल लॉजिक से शुरुआत करें और ज़रूरी हो, तो सुविधाओं और अनुमान के लिए इंटेंशनल लॉजिक की ओर बढ़ें.

ट्रेनिंग के लिए जांचे गए, ज्ञात अच्छे सेट का इस्तेमाल करना डेटा पॉइज़निंग और दूसरे हमलों से बचने का एक अच्छा तरीका है. हालांकि, ज़्यादा अनुमान वाली सुविधाएं बनाने के लिए नया डेटा इंजेस्ट किया जाना चाहिए; ज्ञात अच्छे इनपुट से अपेक्षित नतीजों के आधार पर इस ट्रेनिंग के नतीजों को सावधानी से टेस्ट करने से ज़्यादा बेहतर सिस्टम मिलते हैं.

पक्का करें कि इंटीग्रेशन टेस्टिंग को प्रिमिटिव के इस्तेमाल के इच्छित उदाहरणों के साथ बनाया गया है.

मशीन लर्निंग वर्कफ़्लो को किसी भी दूसरे सॉफ़्टवेयर की तरह इंटीग्रेशन और यूनिट टेस्टिंग की ज़रूरत होती है. एमएल सॉफ़्टवेयर की अनुमानी प्रकृति, विश्वसनीय दावे बनाना मुश्किल बना सकती है. इसलिए, वर्कफ़्लो के इच्छित, लंबे समय के नतीजों पर फ़ोकस करने का सुझाव दिया जाता है.

उन हमलों पर नज़र रखें, जो मशीन सिंथेसिस का इस्तेमाल करते हैं, खास तौर पर उन कंट्रोल पर नज़र रखें जो “what-you-are” पर फ़ोकस करते हैं.

आवाज़, वीडियो, और टेक्स्ट की प्रामाणिकता पर कोई राय बनाने से पहले हमेशा पूछें "क्या कोई मशीन अभी ऐसा कर सकती है?". ऐसा खास तौर पर तब करना ज़रूरी है, जब ऐक्सेस करने के अधिकार देने में इसका इस्तेमाल होता हो.

आर्थिक मॉडल से आगे रहने और एक मजबूत, कम लागत वाला सुरक्षा मॉडल बनाने के लिए म्यूचुअल पॉलिसी का इस्तेमाल करें.

अगर किसी हमले की लागत उसकी बाद की जवाबी कार्रवाई करने की लागत से कम है, तो कंट्रोल अप्रभावी होने की संभावना है. सुरक्षा की लागतों को, सफल हमलों के असर के हिसाब से अपने-आप अडजस्ट होने के लिए ऑपरेशनलाइज़ करें. इससे, धीरे-धीरे हमलावर के लिए हमले की लागत बढ़ती जाएगी.

Google

मालिकाना हक वाले डेटा पर ट्रेन किए गए मॉडल को उपयोगकर्ता के कंट्रोल में रखने से बचें, जैसे किसी डिवाइस पर.

मशीन लर्निंग प्रिमिटिव बनाने में शामिल इस्तेमाल के उदाहरणों के बारे में सोचें और सिद्धांत के तौर पर, मान लें कि इसे बनाने वाले ट्रेनिंग डेटा को फिर से पाया जा सकता है. प्रॉडक्ट डेवलप करते समय कारोबार को जारी रखने की योजना और जोखिम का आकलन करने के लिए इसका इस्तेमाल करें.

एक्सट्रपलेशन हमलों से बचने और संसाधनों की गैर-ज़रूरी खपत को रोकने के लिए कीमतों को सीमित करने की तकनीक का इस्तेमाल करें.

पारंपरिक कंट्रोल अब भी काफ़ी हद तक असरदार हैं. इनमें, मशीन लर्निंग तकनीक को डिज़ाइन और इंप्लीमेंट करते समय कई लेयर में कंट्रोल लागू करना जैसे कि एक्सेस कंट्रोल सूचियां, कीमत सीमित करना और प्रमाणीकरण.