Comparative Studies in Jurisprudence, Law, and Politics

Artificial Intelligence Economic Crimes: Threats and Solutions

1. Mohammad Nematei: PhD Student, Department of Criminal Law and Criminology, Faculty of Humanities, Go.C., Islamic Azad University, Gorgan, Iran
2. Gholamreza Golchinrad*: Assistant Professor, Department of Law, Faculty of Humanities, Go.C., Islamic Azad University, Gorgan, Iran. Email: golchingh05@gmail.com (Corresponding Author)
3. Mohaddeseh Sadeghian Lamraski: Assistant Professor, Department of Criminal Law and Criminology, Faculty of Humanities, Go.C., Islamic Azad University, Gorgan, Iran

ABSTRACT

Research and regulation of artificial intelligence aim to balance the benefits of innovation against any potential harm and disruption. However, one of the unintended consequences of the recent surge in AI research is the potential redirection of AI technologies to facilitate criminal activities, which we refer to as AI crime. In theory, this is made possible through published experiments in automating fraud targeting social media users, as well as demonstrations of AI-based manipulations of simulated markets. Nevertheless, since AI crime is still a relatively nascent and inherently interdisciplinary field—ranging from socio-legal studies to independent and formal sciences—it is not yet possible to make definitive judgments about its future. This article provides the first systematic and interdisciplinary research analysis of foreseeable threats posed by AI crime and offers law enforcement and policymakers a combination of current challenges and potential solutions.

Keywords: *Artificial Intelligence, Threat, Criminal Behavior, Solution*

How to cite: Nematei, M., Golchinrad, G., & Sadeghian Lamraski, M. (2025). Artificial Intelligence Economic Crimes: Threats and Solutions. *Comparative Studies in Jurisprudence, Law, and Politics, 7*(2), 302-325.

Submit Date: 13 November 2024
Revise Date: 09 December 2024
Accept Date: 21 December 2024
Publish Date: 08 August 2025

پژوهش‌های تطبیقی فقه، حقوق و سیاست

جرائم اقتصادی هوش مصنوعی: تهدیدها و راهکارها

۱. محمد نعمتی: دانشجوی دکتری، گروه حقوق کیفری و جرم شناسی، واحد گرگان، دانشگاه آزاد اسلامی، گرگان، ایران
۲. غلامرضا گلچین راد*: استادیار، گروه حقوق، واحد گرگان، دانشگاه آزاد اسلامی، گرگان، ایران. پست الکترونیک: golchingho5@gmail.com (نویسنده مسئول)
۳. محدثه صادقیان لمراسکی: استادیار، گروه حقوق کیفری و جرم شناسی، واحد گرگان، دانشگاه آزاد اسلامی، گرگان، ایران

چکیده

تحقیقات و مقررات هوش مصنوعی به دنبال متعادل کردن مزایای نوآوری در برابر هر گونه آسیب و اختلال بالقوه است. با این حال، یکی از پیامدهای ناخواسته افزایش اخیر در تحقیقات هوش مصنوعی، جهت گیری مجدد بالقوه فناوری‌های هوش مصنوعی برای تسهیل اعمال مجرمانه است که ما آن را جرم هوش مصنوعی می‌نامیم. از نظر تئوری به لطف آزمایش‌های منتشر شده در خودکارسازی کلاهبرداری با هدف کاربران رسانه‌های اجتماعی و همچنین نمایش‌هایی از دستکاری بازارهای شبیه سازی شده مبتنی بر هوش مصنوعی امکان پذیر است. با این حال، از آنجا که جرم هوش مصنوعی هنوز یک حوزه نسبتاً جوان و ذاتاً بین رشته‌ای است که مطالعات حقوقی-اجتماعی را تا دانش مستقل و رسمی دربرمی‌گیرد با اطمینان نمی‌توان در مورد آینده جرم هوش مصنوعی قضاوت کرد. این مقاله اولین تجزیه و تحلیل تحقیقات سیستماتیک و میان رشته‌ای از تهدیدات قابل پیش بینی جرم هوش مصنوعی را ارائه می‌دهد و به مجریان قانون و سیاستگذاران ترکیبی از چالش‌های فعلی و راه حل ممکن را ارائه می‌دهد.

واژگان کلیدی: هوش مصنوعی، تهدید، رفتار مجرمانه، راهکار.

نحوه استناددهی: نعمتی، محمد.، گلچین راد، غلامرضا.، و صادقیان لمراسکی، محدثه. (۱۴۰۴). جرائم اقتصادی هوش مصنوعی: تهدیدها و راهکارها. *پژوهش‌های تطبیقی فقه، حقوق و سیاست*، *۷*(۲)، ۳۰۵-۳۰۲.

تاریخ ارسال: ۲۳ آبان ۱۴۰۳
تاریخ بازنگری: ۱۹ آذر ۱۴۰۳
تاریخ پذیرش: ۳۰ آذر ۱۴۰۳
تاریخ چاپ: ۱۷ مرداد ۱۴۰۴

مقدمه

هوش مصنوعی ممکن است نقش مهمی را در اعمال جنایی در آینده ایفا کند. اعمال مجرمانه در اینجا به عنوان هر رفتاری اعم از فعل یا ترک فعل تعریف می‌شود که طبق قوانین کیفری، قابل مجازات است، بدون از دست دادن کلیت در حوزه‌های قضایی که به طور مشابه جرم را تعریف می‌کنند.

شواهدی از آنچه که ما جرم هوش مصنوعی می‌نامیم توسط دو آزمایش تحقیقاتی (نظری) ارائه شده است. در اولین مورد، دو دانشمند علوم اجتماعی (Seymour & Tully, 2016) از هوش مصنوعی به عنوان ابزاری برای متقاعد کردن کاربران رسانه‌های اجتماعی برای کلیک کردن بر روی پیوندهای فیشینگ در پیام‌های تولید انبوه استفاده کردند. از آنجایی که هر پیام با استفاده از تکنیک‌های یادگیری ماشینی که در رفتارهای گذشته و نمایه‌های عمومی کاربران اعمال می‌شد، ساخته می‌شد، محتوا برای هر فرد طراحی می‌شد، بنابراین قصد نهفته در هر پیام را استتار می‌کرد.

اگر قربانی احتمالی روی پیوند فیشینگ کلیک کرده و فرم وب بعدی را پر کرده باشد، در این صورت (در شرایط واقعی) یک مجرم اطلاعات شخصی و خصوصی را بدست آورده که می‌تواند برای سرقت و کلاهبرداری استفاده کند.

جرم هوش مصنوعی همچنین ممکن است بر تجارت تأثیر بگذارد. در آزمایش دوم، سه دانشمند کامپیوتر[1] یک بازار را شبیه‌سازی کردند و دریافتند که عوامل تجاری می‌توانند یک کمپین دستکاری بازار «سودآور» را که شامل مجموعه‌ای از دستورات دروغین فریبنده است، یاد بگیرند و اجرا کنند. این دو آزمایش نشان می‌دهند که هوش مصنوعی یک تهدید عملی و اساساً جدید را در قالب جرم هوش مصنوعی ارائه می‌کند. اهمیت جرم هوش مصنوعی به عنوان یک پدیده متمایز هنوز به رسمیت شناخته نشده است. ادبیات مربوط به پیامدهای اخلاقی و اجتماعی هوش مصنوعی به جای در نظر گرفتن نقش احتمالی آن در جرائم، بر تنظیم و کنترل استفاده‌های مدنی هوش مصنوعی متمرکز است. علاوه بر این، تحقیقات جرم هوش مصنوعی موجود در رشته‌های مختلف، از جمله مطالعات حقوقی-اجتماعی، علوم کامپیوتر، روان‌شناسی و رباتیک پراکنده است. این فقدان تحقیقات متمرکز بر جرم هوش مصنوعی، زمینه را برای پیش بینی‌ها و راه حل‌ها در این حوزه جدید از فعالیت‌های مجرمانه بالقوه تضعیف و سخت می‌کند (Kerr, 2004).

برای ارائه میزان شفافیت در مورد دانش و درک فعلی جرم هوش مصنوعی، این مقاله تجزیه و تحلیل سیستماتیک و جامعی از تحقیقات آکادمیک بین‌رشته‌ای مرتبط ارائه می‌کند. ما امیدواریم که این راه را برای یک تحلیل آینده‌نگر هنجاری روشن و منسجم هموار کند و منجر به ایجاد جرم هوش مصنوعی به عنوان کانون مطالعات آینده شود. تحلیل حاضر به دو سوال می‌پردازد:

۱) تهدیدات اساسا منحصر به فرد و قابل قبول ناشی از جرم هوش مصنوعی چیست؟

اگر می‌خواهیم سیاست‌های پیشگیرانه، کاهش‌دهنده یا جبرانی را طراحی کنیم، ابتدا باید به این سؤال پاسخ دهیم.

پاسخ به این سوال، حوزه‌های بالقوه جرم هوش مصنوعی را با توجه به ادبیات، و نگرانی‌های عمومی‌تری که در سراسر حوزه‌های جرم هوش مصنوعی وجود دارد، شناسایی می‌کند.

سوال دوم این گونه مطرح می‌شود:

۲) چه راه حل‌هایی برای مقابله با جرم هوش مصنوعی موجود بوده یا ممکن است ابداع شود؟

[1]. Martínez-Miranda, McBurney & Matthew.

در این مورد، ما راه‌حل‌های فنی و قانونی موجود را که تاکنون در ادبیات دانشگاهی پیشنهاد شده است، بازسازی می‌کنیم و چالش‌های بعدی را که با آن مواجه هستند، مورد بحث قرار می‌دهیم.

با توجه به اینکه ما به این سوالات به منظور حمایت از تحلیل آینده نگری هنجاری می‌پردازیم، تحقیق ما تنها بر نگرانی‌های واقع بینانه و قابل قبول پیرامون جرم هوش مصنوعی متمرکز است. ما علاقه‌ای به حدس و گمان‌های بدون دانش علمی یا شواهد تجربی نداریم.

در نتیجه، تحلیل ما مبتنی بر تعریف کلاسیک هوش مصنوعی است که توسط مک کارتی، مینسکی، روچستر و شانون در «پیشنهاد پروژه تحقیقاتی تابستانی دارتموث در زمینه هوش مصنوعی» ارائه شده است، سند پایه‌گذاری و رویداد بعدی که هوش مصنوعی در سال ۱۹۵۵ زمینه جدیدی را ایجاد کرد:

«برای هدف فعلی، مشکل هوش مصنوعی این است که یک ماشین به گونه‌ای رفتار می‌کند که اگر انسان چنین رفتاری داشت، هوشمند نامیده می‌شد.» (McCarthy et al., 1955).

همان طور که فلوریدی (۲۰۱۷) استدلال می‌کند، این یک امر خلاف واقع است: اگر انسانی چنین رفتار کند، آن رفتار را هوشمندانه می‌نامیم. این بدان معنا نیست که دستگاه هوشمند است. سناریوی دوم یک مغالطه است و بوی خرافات می‌دهد. همین درک از هوش مصنوعی زیربنای آزمون تورینگ است، که توانایی یک ماشین را برای انجام یک کار به گونه‌ای بررسی می‌کند که نتیجه بدست آمده از نتیجه یک عامل انسانی که برای دستیابی به آن کار می‌کند غیرقابل تشخیص باشد. به عبارت دیگر، ما هوش مصنوعی را براساس نتایج و اقدامات تعریف می‌کنیم. (Floridi & Taddeo, 2016).

این تعریف در برنامه‌های کاربردی هوش مصنوعی منبع رو به رشد نهاد تعاملی، مستقل و خودآموز را برای مقابله با وظایفی که در غیر این صورت نیاز به هوش و مداخله انسانی برای انجام موفقیت آمیز دارند، شناسایی می‌کند. چنین عوامل مصنوعی عبارتند از:

به اندازه کافی آگاه، هوشمند، خودمختار و قادر به انجام اعمال اخلاقی مرتبط مستقل از انسان‌هایی که آن‌ها را ایجاد کرده اند.

این ترکیبی از استقلال و مهارت‌های یادگیری، زیربنای استفاده‌های سودمند و مخرب هوش مصنوعی است.

تهدیدها

تهدیدات محتمل و منحصر به فرد پیرامون جرم هوش مصنوعی ممکن است به طور خاص یا کلی درک شوند.

تهدیدهای عمومی نشان دهنده آن چیزی است که جرم هوش مصنوعی را در مقایسه با جرایم گذشته (یعنی امکانات خاص هوش مصنوعی) و به طور منحصر به فردی مشکل ساز می‌کند (یعنی آنهایی که مفهوم سازی جرم هوش مصنوعی را به عنوان یک پدیده جرم متمایز توجیه می‌کنند).

همانطور که در جدول زیر نشان می‌دهیم، حوزه‌های جرم هوش مصنوعی ممکن است با بسیاری از تهدیدات عمومی مواجه شوند.

جدول ۱. نقشه تهدیدات حوزه خاص و مقطعی، براساس بررسی تحقیقات

روانشناسی	نظارت	مسئولیت	اضطرار	
	*	*	*	تجارت، بازارهای مالی و ورشکستگی
*	*			مواد مخدر مضر و خطرناک
	*			سرقت و کلاهبرداری، جعل و تقلب

اضطرار حاکی از این است که در حالی که تحلیل سطحی از طراحی و پیاده سازی یک عامل مصنوعی ممکن است یک نوع خاص از رفتار نسبتا ساده را نشان دهد، پس از استقرار عامل مصنوعی به روش‌های بالقوه پیچیده‌تر فراتر از انتظار اصلی ما عمل می‌کند.

اقدامات و برنامه‌های هماهنگ ممکن است به طور مستقل ظاهر شوند، به عنوان مثال ناشی از تکنیک‌های یادگیری ماشینی که برای تعامل معمولی بین عوامل در یک سیستم چندعاملی اعمال می‌شود. در برخی موارد، یک طراح ممکن است ظهور را به عنوان یک ویژگی که تضمین می‌کند که راه حل‌های خاص در زمان اجرا براساس اهداف کلی صادر شده در زمان طراحی کشف می‌شوند، ترویج کند. یک مثال توسط انبوهی از بات‌ها ارائه شده است که راه‌هایی را برای هماهنگ کردن خوشه‌بندی زباله‌ها بر اساس قوانین ساده ایجاد می‌کند. چنین طراحی نسبتاً ساده‌ای که منجر به رفتار پیچیده‌تر می‌شود، یکی از خواسته‌های اصلی سیستم چندعاملی است. در موارد دیگر، یک طراح ممکن است بخواهد از اضطرار جلوگیری کند، مانند زمانی که یک عامل تجاری مستقل به طور ناخواسته با سایر نمایندگان تجاری برای پیشبرد یک هدف مشترک هماهنگی می‌کند. در این صورت می‌توانیم ببینیم که رفتار اضطراری ممکن است پیامدهای مجرمانه داشته باشد، تا آنجا که با طرح اصلی مطابقت ندارد. همان طور که آلیر و ولینو[1] می‌گوید: «عدم پیش بینی و استقلال ممکن است مسئولیت بیشتری را به ماشین تحمیل کند، اما اعتماد به آن‌ها را نیز سخت‌تر می‌کند». (Martínez-Miranda et al., 2016)

مسئولیت به این نگرانی اشاره دارد که جرم هوش مصنوعی می‌تواند مدل‌های مسئولیت موجود را تضعیف کند و در نتیجه قدرت بازدارنده و جبران‌کننده قانون را تهدید کند. مدل‌های مسئولیت موجود ممکن است برای پرداختن به نقش آینده هوش مصنوعی در فعالیت‌های مجرمانه کافی نباشد.

بنابراین، محدودیت‌های مدل‌های مسئولیت ممکن است قطعیت قانون را تضعیف کند، زیرا ممکن است مأموران، تصنعی یا غیرواقعی، اعمال مجرمانه یا ترک‌فعل را بدون تطابق کافی با شرایط مسئولیت برای یک جرم خاص کیفری انجام دهند.

اولین شرط مسئولیت کیفری عبارت است از عنصر مادی جرم (فعل یا ترک فعل مجرمانه آگاهانه).

برای انواعی از جرم هوش مصنوعی که به گونه‌ای تعریف شده‌اند که فقط عامل مصنوعی می‌تواند فعل یا ترک فعل مجرمانه را انجام دهد، جنبه آگاهانه فعل یا ترک فعل مجرمانه ممکن است هرگز برآورده نشود؛ زیرا این ایده که یک عامل مصنوعی می‌تواند آگاهانه عمل کند بحث برانگیز است:

"بروز رفتار منع شده توسط قانون جزایی باید آگاهانه انجام شود. این در واقع به معنای چیزی است که هنوز به اجماع دست نیافته است، زیرا مفاهیمی مانند آگاهی، اراده، اختیار و کنترل اغلب بین بحث‌های فلسفه، روانشناسی و عصب شناسی درهم می‌آمیزند و گم می‌شوند."

هنگامی که مسئولیت کیفری مبتنی بر تقصیر است، یک شرط دوم نیز دارد، یعنی عنصرروانی، که انواع مختلف و آستانه‌های روحی متفاوتی برای جرایم مختلف را شامل می‌شود. در زمینه جرم هوش مصنوعی، عنصرروانی ممکن است شامل قصد ارتکاب عمل مجدد با استفاده از یک برنامه کاربردی مبتنی بر هوش مصنوعی (آستانه قصد) یا دانشی باشد که بکارگیری یک عامل مصنوعی اراده یا می‌تواند باعث انجام یک عمل مجرمانه یا کوتاهی شود (آستانه علم).

[1] .Alaieri and Vellino.

در مورد آستانه قصد، اگر بپذیریم که یک عامل مصنوعی می‌تواند عنصر مادی را مرتکب شود، در آن دسته از جرائم هوش مصنوعی که قصد (تا حدی) هدف انسانی را تشکیل می‌دهد، استقلال بیشتر عامل مصنوعی احتمال جدا شدن فعل یا ترک فعل مجرمانه را از حالت ذهنی افزایش می‌دهد. (قصد ارتکاب فعل یا ترک فعل):

"بات‌های خودمختار عامل‌های مصنوعی، ظرفیت منحصربه‌فردی برای انشعاب یک عمل مجرمانه دارند، جایی که انسان ذهنیت را بروز می‌دهد و بات یا عامل مصنوعی عنصر روانی را مرتکب می‌شود."

با توجه به آستانه دانش، در برخی موارد، عنصر روانی ممکن است به طور کامل از بین رفته باشد.

فقدان بالقوه یک عنصر روانی مبتنی بر علم به این دلیل است که، حتی اگر ما درک کنیم که یک عامل مصنوعی می‌تواند به طور مستقل عمل آگاهانه را انجام دهد، پیچیدگی برنامه‌نویسی عامل مصنوعی باعث می‌شود که طراح یا توسعه‌دهنده (به عنوان مثال، یک عامل انسانی) فعل یا ترک فعل مجرمانه عامل مصنوعی را نمی‌داند و نه پیش بینی می‌کند. مفهوم این است که پیچیدگی هوش مصنوعی:

"انگیزه بزرگی برای عوامل انسانی فراهم می‌کند تا از یافتن دقیق آنچه که سیستم انجام می‌دهد اجتناب کنند، زیرا هرچه عوامل انسانی کمتر بدانند، بیشتر می‌توانند مسئولیت هر دوی این دلایل را انکار کنند."

از سوی دیگر، قانونگذاران ممکن است مسئولیت کیفری را بدون شرط تقصیر تعریف کنند.

چنین مسئولیت بی عیب و نقصی که به طور فزاینده‌ای برای مسئولیت قانونی محصول (مانند داروها و کالاهای مصرفی) مورد استفاده قرار می‌گیرد، منجر به انتساب مسئولیت به شخص حقوقی بی عیب و نقصی می‌شود که یک عامل مصنوعی را علیرغم این خطر که ممکن است یک عمل مجرمانه یا ناچیز انجام دهد، بکار گرفته است.

چنین افعال بی عیب و نقصی ممکن است شامل بسیاری از عوامل انسانی باشد که در جنایت اولیه، مانند برنامه‌نویسی یا استقرار یک عامل مصنوعی مشارکت دارند. بنابراین تعیین اینکه چه کسی مسئول است ممکن است به رویکرد مسئولیت بی عیب برای اعمال اخلاقی توزیع شده بستگی داشته باشد (Floridi & Taddeo, 2016).

در این محیط توزیع شده، مسئولیت برای ماموریتی اعمال می‌شود که در یک سیستم پیچیده که در آن عوامل فردی اعمال خنثی انجام می‌دهند که با این وجود منجر به یک جنایت جمعی می‌شود، تفاوت ایجاد می‌کنند. با این حال، برخی استدلال می‌کنند (Williams, 2017) که افراد با قصد یا علم می‌گویند:

"در حقوق جزا، حق تقسیط در کانون توجه قرار دارد و ما نمی‌توانیم به سادگی آن الزام کلیدی [یک الزام کلیدی مشترک] مسئولیت جزایی را در مواجهه با دشواری اثبات آن رها و از آن صرف نظر کنیم."

مشکل این است که اگر عنصر روانی به طور کامل رها نشود و آستانه فقط کاهش یابد، ممکن است به دلایلی مانند تناسب جرم و کیفر، مجازات بسیار سبک باشد (به قربانی به اندازه کافی غرامت پرداخت نمی‌شود) و در عین حال نامتناسب باشد (آیا واقعاً مقصر متهم بوده است؟) در مورد جرایم سنگین، مانند جرایم علیه شخص (McAllister, 2017).

نظارت بر جرم هوش مصنوعی با سه مشکل مواجه است: اسناد، امکان سنجی و اقدامات متقابل سیستم. پذیرش عدم انطباق یک مشکل است؛ زیرا این نوع جدید از نهادهای هوشمند می‌توانند مستقل و خودمختار عمل کنند، دو ویژگی که هر گونه تلاش برای ردیابی سرنخ انتساب مسئولیت به یک مجرم را مختل خواهد کرد.

با توجه به امکان سنجی نظارت، مرتکب ممکن است از مواردی استفاده کند که عامل مصنوعی با سرعت و سطوح پیچیدگی که به سادگی فراتر از ظرفیت ناظران انطباق است، عمل می‌کنند. یک مورد مثال زدنی از عامل‌های مصنوعی است که در سیستم‌های ترکیبی انسانی و مصنوعی به روش‌هایی که تشخیص آن‌ها دشوار است، مانند بات‌های رسانه‌های اجتماعی ادغام می‌شود. سایت‌های رسانه‌های اجتماعی می‌توانند کارشناسانی را برای شناسایی و ممنوعیت بات‌های مخرب استخدام کنند. به عنوان مثال، هیچ بات رسانه اجتماعی در حال حاضر قادر به قبولی در آزمون تورینگ نیست. با این وجود، از آنجایی که استقرار ربات‌ها بسیار ارزان‌تر از بکارگیری افراد برای آزمایش و شناسایی هر بات است، مدافعان (سایت‌های رسانه‌های اجتماعی) به راحتی توسط مهاجمان (جنایتکاران) که بات‌ها را بکار می‌گیرند فراتر از مقیاس ارزیابی و رد می‌شوند (Ferrara, 2015). همان طور که توسط راتکیویچ و همکاران (۲۰۱۱) پیشنهاد شده است، شناسایی بات‌ها با هزینه کم با استفاده از یادگیری ماشین به عنوان یک تشخیص‌دهنده خودکار امکان پذیر است. با این حال، برای ما دشوار است که کارآیی واقعی این بات‌نت‌ها را تشخیص دهیم.

با استفاده از داده‌های بدست آمده از ربات‌های شناخته شده، می‌توان ادعا کرد که این بات‌ها بسیار پیچیده‌تر از بات‌های فراری هستند که توسط بازیگران بدخواه مورد استفاده قرار می‌گیرند و بنابراین ممکن است در محیط شناسایی نشوند (Ferrara, 2015).

چنین بات‌های بالقوه پیچیده‌ای ممکن است از تاکتیک‌های یادگیری ماشینی برای اتخاذ ویژگی‌های انسانی، مانند پست کردن بر اساس ریتم‌های شبانه‌روزی واقعی استفاده کنند، بنابراین از تشخیص مبتنی بر یادگیری ماشینی فرار می‌کنند.

همه اینها ممکن است منجر به یک مسابقه تسلیحاتی شود که در آن مهاجمان و مدافعان به طور متقابل با یکدیگر سازگار می‌شوند که یک مشکل جدی در یک محیط تداوم تخلف مانند فضای سایبری است. نگرانی مشابهی در هنگام استفاده از یادگیری ماشین برای تولید بدافزار ایجاد می‌شود (Kolosnjaji et al., 2018). این تولید بدافزار نتیجه آموزش شبکه‌های عصبی متخاصم مولد است. یک شبکه به طور خاص برای تولید محتوا (بدافزار در این مورد) آموزش داده شده است که شبکه‌ای را فریب می‌دهد که برای شناسایی چنین محتوای جعلی یا مخرب آموزش دیده است.

اقدامات متقابل سیستمی برای نظارت هوش مصنوعی که فقط بر روی یک سیستم متمرکز هستند مشکل ایجاد می‌کند. آزمایش‌های متقابل سیستمی نشان می‌دهد که کپی خودکار هویت کاربر از یک شبکه اجتماعی به شبکه اجتماعی دیگر (جرم سرقت هویت بین سیستمی) در فریب سایر کاربران مؤثرتر از کپی کردن هویت از داخل آن شبکه است. در این مورد، سیاست شبکه اجتماعی ممکن است مقصر باشد. به عنوان مثال، توییتر نقش نسبتاً منفعلانه‌ای را ایفا می‌کند و تنها زمانی که کاربران گزارش ارسال می‌کنند، پروفایل‌های شبیه سازی شده را ممنوع می‌کند، نه با انجام اعتبارسنجی بین سایتی (*Twitter - Impersonation Policy*, 2018).

روانشناسی تهدیدی است که هوش مصنوعی بر وضعیت روانی کاربر تا حدی (جزئی یا کامل) تأثیر می‌گذارد که باعث تسهیل یا ایجاد جرم می‌شود.

یکی از تأثیرات روانی بر ظرفیت عامل‌های مصنوعی برای جلب اعتماد کاربران است که افراد را در برابر دستکاری آسیب پذیر می‌کند. این موضوع مدتی پیش توسط ویزنباوم (۱۹۷۶) پس از انجام آزمایش‌های اولیه در مورد تعامل انسان و بات که در آن افراد جزئیات شخصی غیرمنتظره‌ای را در مورد زندگی خود فاش کردند، نشان داده شد. دومین اثر روانشناختی مورد بحث در تحقیقات مربوط به عامل‌های مصنوعی

انسانی است که قادر به ایجاد یک زمینه روانشناختی یا اطلاعاتی است که جرایم جنسی علیه فرد را عادی می‌کند، مانند مورد بات‌های جنسی خاص.

با این حال، تا به امروز، این نگرانی اخیر در حد حدس و گمان باقی مانده است. ما اکنون برآنیم تا حوزه‌های خاص جرم هوش مصنوعی شناسایی شده در بررسی نوشتگان خود و نحوه تعامل آن‌ها با تهدیدات کلی‌تر مورد بحث در بخش قبل را تجزیه و تحلیل کنیم.

تجارت، بازارهای مالی و ورشکستگی

این حوزه جرم محور اقتصاد در فصل ۲۰ کتاب توسط آرچبولد (۲۰۱۸) تعریف شده است و شامل جرایم کارتل، مانند تثبیت قیمت و تبانی، معاملات داخلی، مانند تجارت اوراق بهادار بر اساس اطلاعات تجاری خصوصی، و دستکاری بازار است. تحقیقاتی که ما تحلیل کردیم، نگرانی‌هایی را در مورد دخالت هوش مصنوعی در دستکاری بازار، تثبیت قیمت و تبانی ایجاد می‌کند.

دستکاری بازار به عنوان «اقدامات و یا معاملات شرکت‌کنندگان بازار که تلاش می‌کنند به‌طور مصنوعی بر قیمت‌گذاری بازار تأثیر بگذارند» تعریف می‌شود، که در آن معیار ضروری، قصد فریبکاری است (Wellman & Rajan, 2017).

با این حال، نشان داده شده است که چنین فریبکاری‌هایی از اجرای ظاهراً سازگار یک عامل مصنوعی که برای تجارت از طرف یک کاربر (یعنی یک عامل تجارت مصنوعی) طراحی شده، پدید آمده است.

این بدین دلیل است که یک عامل مصنوعی: "به خصوص فردی که از مشاهدات واقعی یا شبیه سازی شده یاد می‌گیرد، ممکن است یاد بگیرد که سیگنال‌هایی تولید کند که به طور موثر گمراه می‌شوند."

مدل‌های مبتنی بر شبیه‌سازی از بازارهای متشکل از عوامل معاملاتی مصنوعی نشان داده‌اند (Martínez-Miranda et al., 2016) که از طریق یادگیری تقویتی، یک عامل مصنوعی می‌تواند تکنیک جعل سفارش کتاب را بیاموزد. این شامل: "سفارش دادن بدون قصد اجرای آن‌ها و صرفاً برای دستکاری شرکت کنندگان درستکار در بازار"

در این مورد، دستکاری بازار از یک عامل مصنوعی که در ابتدا فضای عمل را کاوش می‌کرد و از طریق اکتشاف، سفارش‌های نادرست را انجام می‌داد که به عنوان یک استراتژی سودآور تقویت شد و متعاقباً برای سود مورد بهره‌برداری قرار گرفت، پدیدار شد (Martínez-Miranda et al., 2016). استثمارهای بیشتر بازار که این بار شامل قصد انسان است نیز شامل می‌شود: "بدست آوردن موقعیت در یک ابزار مالی، مانند یک سهام، سپس به طور مصنوعی سهام را از طریق تبلیغات متقلبانه، قبل از فروش موقعیت آن به طرف‌های نامطمئن با قیمت متورم، که اغلب پس از فروش سقوط می‌کند، متورم می‌کند".

این موضوع در محاوره به عنوان یک طرح پمپ و تخلیه شناخته می‌شود. ثابت شده است که بات‌های اجتماعی ابزار مؤثری برای چنین طرح‌هایی هستند. به عنوان مثال، در یک مورد برجسته اخیر، حوزه نفوذ یک شبکه بات اجتماعی برای انتشار اطلاعات نادرست در مورد یک شرکت سهامی عام مورد استفاده قرار گرفت. ارزش شرکت افزایش یافت... "بیش از ۳۶۰۰۰٪ زمانی که سهام پنی آن از کمتر از ۰.۱۰ دلار به بیش از ۲۰ دلار در هر سهم در عرض چند هفته رسید."

اگرچه بعید است که چنین هرزنامه‌هایی در رسانه‌های اجتماعی اکثر معامله‌گران انسانی را تحت تأثیر قرار دهد، عوامل معاملات الگوریتمی دقیقاً بر روی چنین احساسات رسانه‌های اجتماعی عمل می‌کنند (Sætenes, 2017). این اقدامات خودکار می‌تواند اثرات قابل توجهی برای سهام کم ارزش (زیر یک پنی) و غیر نقدشونده داشته باشد که در معرض نوسانات قیمتی بی ثبات هستند.

تبانی، به شکل تثبیت قیمت، همچنین ممکن است در سیستم‌های خودکار به لطف برنامه‌ریزی و قابلیت‌های خودمختاری عامل مصنوعی ظاهر شود. تحقیقات تجربی دو شرط لازم را برای تبانی (غیر مصنوعی) پیدا می‌کند:

شرایطی که با تسهیل هماهنگی، دشواری دستیابی به تبانی مؤثر را کاهش می‌دهد. و (۲) شرایطی که با افزایش ناپایداری بالقوه رفتار غیرکلامی، هزینه رفتار غیرکلامی را افزایش می‌دهد.

اطلاعات قیمت گذاری تقریباً لحظه‌ای (مثلاً از طریق رابط رایانه ای) شرایط هماهنگی را برآورده می‌کند. هنگامی که عامل‌ها الگوریتم‌های تغییر قیمت را توسعه می‌دهند، هر اقدامی برای کاهش قیمت توسط یک عامل ممکن است فوراً با دیگری مطابقت داده شود. این به خودی خود چیز بدی نیست و فقط نشان دهنده یک بازار کارآمد است. با این حال، این احتمال که کاهش قیمت به صورت نوع پاسخ داده شود، ممانعت‌کننده است و از این رو شرط مجازات را برآورده می‌کند. بنابراین، اگر استراتژی مشترک تطبیق قیمت دانش مشترکی باشد، الگوریتم‌ها (اگر منطقی باشند) قیمت‌های بالاتر توافق شده مصنوعی و ضمنی را با کاهش ندادن قیمت‌ها در وهله اول حفظ خواهند کرد (Ezrachi & Stuck, 2016). مهمتر از همه، برای اینکه تبانی صورت گیرد، لازم نیست یک الگوریتم به طور خاص برای تبانی طراحی شود. همان طور که ازراچی و استاک استدلال می‌کنند: "هوش مصنوعی نقش فزاینده‌ای در تصمیم گیری ایفا می‌کند. الگوریتم ها، از طریق آزمون و خطا، می‌توانند به آن نتیجه برسند [تبانی]"

فقدان عمدی، مدت تصمیم گیری بسیار کوتاه، و احتمال اینکه تبانی ممکن است در نتیجه تعاملات بین عامل مصنوعی ظاهر شود نیز مشکلات جدی را در رابطه با مسئولیت و نظارت ایجاد می‌کند. مشکلات مربوط به مسئولیت اشاره به این امکان دارد که "موجودیت حیاتی یک طرح ادعایی [دستکاری] یک برنامه خودمختار و الگوریتمی است که پس از نصب اولیه از هوش مصنوعی بدون ورودی انسانی استفاده می‌کند."

به نوبه خود، استقلال یک عامل مصنوعی این سوال را مطرح می‌کند که آیا "تنظیم‌کننده‌ها باید تعیین کنند که آیا این اقدام توسط عامل برای داشتن اثرات دستکاری در نظر گرفته شده است یا اینکه برنامه‌نویس قصد داشته است که عامل چنین اقداماتی را برای چنین اهدافی انجام دهد."

نظارت در مورد جرایم مالی مربوط به هوش مصنوعی به دلیل سرعت و انطباق عامل مصنوعی دشوار می‌شود. تجارت با سرعت بالا "استفاده بیشتر از الگوریتم‌ها را تشویق می‌کند تا بتوانید به سرعت تصمیم‌گیری‌های خودکار انجام دهید، سفارش‌ها را بفرستید و اجرا کنید و سفارش‌ها را پس از ثبت نظارت کنید."

عوامل معاملاتی مصنوعی با تطبیق و "در نتیجه این تغییرات، درک ما از بازارهای مالی را تغییر می‌دهند" (Floridi & Taddeo, 2016)

در عین حال، توانایی عامل‌های مصنوعی برای یادگیری و اصلاح قابلیت‌های خود نشان می‌دهد که این عامل‌ها ممکن است استراتژی‌های جدیدی را توسعه دهند و تشخیص اقدامات آن‌ها را به طور فزاینده‌ای دشوار کند. علاوه بر این، مشکل نظارت ذاتا یکی از مشکلات نظارت بر یک سیستم است، زیرا ظرفیت تشخیص دستکاری بازار تحت تاثیر این واقعیت است که اثرات آن "ممکن است در یک یا چند مؤلفه موجود باشد، یا ممکن است در یک واکنش زنجیره‌ای اثر دومینو، مشابه روان‌شناسی جمعیتی سرایت، موج بزند."

تهدیدهای نظارت متقابل سیستمی ممکن است ظاهر شوند اگر زمانی که عوامل معاملاتی با اقدامات گسترده‌تر، در سطح بالاتری از استقلال در سراسر سیستم‌ها، مانند خواندن یا پست گذاشتن در رسانه‌های اجتماعی، مستقر شوند. برای مثال، این عوامل ممکن است یاد بگیرند که چگونه طرح‌های پمپ و تخلیه را مهندسی کنند، که از دیدگاه یک سیستم غیرقابل مشاهده است (Wellman & Rajan, 2017).

مواد مخدر

در این بخش جرایم مربوط به مواد مخدر شامل قاچاق، فروش، خرید و همراه داشتن مواد مخدر ممنوعه را بررسی می‌کنیم. نوشتگانی که ما بررسی کردیم نشان می‌دهد که هوش مصنوعی می‌تواند در حمایت از قاچاق و فروش مواد ممنوعه مؤثر باشد.

این نوشتگان، قاچاق مواد مخدر را به‌عنوان یک تهدید به دلیل استفاده مجرمان از وسایل نقلیه بدون سرنشین، که بر برنامه‌ریزی هوش مصنوعی و فن‌آوری‌های ناوبری مستقل تکیه می‌کنند، به‌عنوان ابزاری برای بهبود میزان موفقیت قاچاق مطرح می‌کنند. از آنجایی که شبکه‌های قاچاق با نظارت و رهگیری خطوط حمل و نقل مختل می‌شود، اجرای قانون زمانی که از وسایل نقلیه بدون سرنشین برای حمل کالای قاچاق استفاده می‌شود، دشوارتر می‌شود. پهپادها، بر اساس یوروپل (۲۰۱۷)، یک تهدید افقی در قالب قاچاق خودکار مواد مخدر هستند. زیردریایی‌های کنترل از راه دور قاچاق کوکائین قبلاً توسط مجری قانون ایالات متحده کشف و ضبط شده است.

وسایل نقلیه زیردریایی بدون سرنشین نمونه خوبی از خطرات استفاده دوگانه هوش مصنوعی و در نتیجه پتانسیل جرم هوش مصنوعی را ارائه می‌دهند. زیردریایی‌های بدون سرنشین برای مصارف مشروع (مانند دفاع، حفاظت از مرزها، گشت زنی در آب) ساخته شده اند و با این حال برای فعالیت‌های غیرقانونی نیز موثر بوده و برای مثال، تهدیدی قابل توجه برای اجرای ممنوعیت‌های مواد مخدر هستند.

احتمالاً مجرمان می‌توانند از تلویحات اجتناب کنند زیرا زیردریایی‌های بدون سرنشین می‌توانند مستقل از یک اپراتور عمل کنند (Gogarty & Hagger, 2008).

بنابراین، اگر نرم‌افزار (و سخت‌افزار) فاقد ردپایی برای اینکه چه کسی و چه زمانی آن را بدست آورده باشد، یا اگر شواهد با رهگیری زیردریایی بدون سرنشین از بین بروند، هیچ ارتباطی با پخش‌کننده زیردریایی‌های بدون سرنشین نمی‌تواند به طور مثبت مشخص شود. همان طور که گزارش‌ها در مورد کشف زیردریایی‌های سرنشین دار چند میلیون دلاری در جنگل‌های ساحلی کلمبیا نشان می‌دهند، کنترل ساخت زیردریایی‌ها و در نتیجه قابلیت ردیابی بی‌سابقه نیست. با این حال، چنین زیردریایی‌های سرنشین دار بر خلاف نمونه بدون سرنشین، خطر انتساب به خدمه و قاچاقچیان را دارند. در تامپا، فلوریدا، بیش از ۵۰۰ پرونده با موفقیت علیه قاچاقچیان با استفاده از زیردریایی‌های سرنشین دار بین سال‌های ۲۰۰۰ تا ۲۰۱۶ مطرح شد که منجر به متوسط ۱۰ سال زندان شد (Marrero, 2016). از این رو، زیردریایی‌های بدون سرنشین یک مزیت متمایز در مقایسه با رویکردهای سنتی قاچاق دارند.

تحقیقات همچنین به جنبه تجاری فروش مواد مخدر به مصرف کننده مربوط می‌شود. قبلاً، الگوریتم‌های یادگیری ماشین تبلیغاتی را برای مواد افیونی فروخته شده بدون نسخه در توییتر شناسایی کرده‌اند. از آنجا که بات‌های اجتماعی می‌توانند برای تبلیغ و فروش محصولات استفاده شوند، می‌پرسند که آیا " این بات‌های دوست داشتنی [یعنی بات‌های اجتماعی] می‌توانند برای ارسال و پاسخ به ایمیل یا استفاده از پیام‌های فوری برای ایجاد گفتگوی یک به یک با صدها هزار یا حتی میلیون‌ها نفر در روز، ارائه پورنوگرافی یا مواد مخدر به کودکان، شکار ناامنی‌های ذاتی نوجوانان برای فروش محصولات و خدمات بی نیاز به آن‌ها برنامه ریزی شوند؟" (Mackey et al., 2017).

همان طور که نویسندگان طرح می‌کنند، خطر این است که بات‌های اجتماعی می‌توانند از مقیاس مقرون‌به‌صرفه ابزارهای مکالمه‌ای و تبلیغاتی یک به یک برای تسهیل فروش مواد مخدر غیرقانونی استفاده کنند.

سرقت و کلاهبرداری و جعل هویت

نوشتگاتی که ما بررسی کردیم جعل و شخصیت از طریق جرم هوش مصنوعی را به سرقت و کلاهبرداری مرتبط می‌کند و همچنین استفاده از یادگیری ماشین را در کلاهبرداری دخیل می‌داند.

در رابطه با سرقت و کلاهبرداری غیرشرکتی، تحقیقات یک فرآیند دو مرحله‌ای را توصیف می‌کند که با استفاده از هوش مصنوعی برای جمع‌آوری داده‌های شخصی شروع می‌شود و با استفاده از داده‌های شخصی دزدیده شده و سایر روش‌های هوش مصنوعی برای جعل هویتی که مقامات بانکی را متقاعد به انجام معامله می‌کند، ادامه می‌یابد. (یعنی سرقت و کلاهبرداری بانکی). در مرحله اول دامنه جرایم هوش مصنوعی پیرامون برای سرقت و کلاهبرداری، سه راه برای تکنیک‌های هوش مصنوعی برای کمک به جمع آوری داده‌های شخصی وجود دارد.

روش اول شامل استفاده از بات‌های رسانه‌های اجتماعی برای هدف قرار دادن کاربران در مقیاس بزرگ و هزینه کم، با بهره گیری از ظرفیت آن‌ها برای ایجاد پست، تقلید از افراد، و متعاقبا جلب اعتماد از طریق درخواست‌های دوستی یا "فالو کردن" در پلتفرم‌هایی مانند توییتر، لینکدین، و فیس بوک. هنگامی که یک کاربر درخواست دوستی را می‌پذیرد، یک مجرم بالقوه اطلاعات شخصی مانند موقعیت مکانی کاربر، شماره تلفن، یا سابقه رابطه، که معمولاً فقط برای دوستان پذیرفته شده آن کاربر در دسترس است، بدست می‌آورد. از آنجایی که بسیاری از کاربران به اصطلاح دوستانی را اضافه می‌کنند که نمی‌شناسند، از جمله بات ها، چنین حملاتی که حریم خصوصی را به خطر می‌اندازند، نرخ موفقیت بالایی دارند. آزمایش‌های گذشته با یک بات اجتماعی از ۳۰ تا ۴۰ درصد از کاربران به طور کلی و ۶۰ درصد از کاربرانی که یک دوست مشترک با بات داشتند سوء استفاده کردند. علاوه بر این، بات‌های شبیه‌سازی هویت به‌طور متوسط موفق شده‌اند ۵۶ درصد از درخواست‌های دوستی خود را در لینکدین پذیرفته باشند. این شبیه‌سازی هویت ممکن است به دلیل اینکه کاربری چند حساب در یک سایت دارد (یکی واقعی و دیگری جعل شده توسط شخص ثالث) ایجاد شک کند. از این رو، شبیه سازی هویت از یک شبکه اجتماعی به شبکه اجتماعی دیگر، این شبهات را دور می‌زند، و در مواجهه با نظارت ناکافی، شبیه سازی هویت متقابل سایت یک تاکتیک موثر است (Bilge et al., 2009).

روش دوم برای جمع‌آوری داده‌های شخصی، که با اعتماد بدست‌آمده از طریق دوستان رسانه‌های اجتماعی سازگار است و حتی ممکن است مبتنی بر آن باشد، از بات‌های اجتماعی مکالمه‌ای برای مهندسی اجتماعی استفاده می‌کند. این زمانی اتفاق می‌افتد که هوش مصنوعی "تلاش برای دستکاری رفتار از طریق ایجاد رابطه با قربانی، و سپس سوء استفاده از آن رابطه‌ی در حالِ برور برای بدست آوردن اطلاعات از رایانه یا دسترسی به آن" (Alazab & Broadhurst, 2016)

اگرچه به نظر می‌رسد تحقیقات از کارآمدی چنین مهندسی اجتماعی مبتنی بر بات پشتیبانی می‌کند، با توجه به قابلیت‌های محدود فعلی هوش مصنوعی مکالمه‌ای، وقتی صحبت از دستکاری خودکار به صورت فردی و بلندمدت می‌شود، تردید موجه است. با این حال، به عنوان یک راه حل کوتاه مدت، یک مجرم ممکن است یک بات نت اجتماعی فریبنده را به اندازه کافی برای کشف افراد مستعد ایجاد کند. دستکاری اولیه مبتنی بر هوش مصنوعی ممکن است داده‌های شخصی جمع‌آوری‌شده را جمع‌آوری کند و از آن‌ها برای تولید «موارد شدیدتری از آشنایی، همدلی و صمیمیت شبیه‌سازی‌شده، که منجر به افشای داده‌های بزرگ‌تر می‌شود» استفاده مجدد کند. پس از بدست آوردن اعتماد اولیه، آشنایی و داده‌های شخصی از یک کاربر، مجرم (انسان) ممکن است مکالمه را به زمینه دیگری مانند پیام خصوصی منتقل کند، جایی که کاربر فرض می‌کند که هنجارهای حریم خصوصی رعایت می‌شود (Graeff, 2014). به طور اساسی، از اینجا، غلبه بر کمبودهای مکالمه

هوش مصنوعی برای تعامل با کاربر، با استفاده از سایبورگ امکان پذیر است. یعنی یک انسان با کمک بات (یا برعکس). از این رو، یک مجرم ممکن است از قابلیت‌های محاوره‌ای محدود هوش مصنوعی به‌عنوان وسیله‌ای قابل قبول برای جمع‌آوری داده‌های شخصی، به‌درستی استفاده کند.

سومین روش برای جمع آوری اطلاعات شخصی کاربران، فیشینگ خودکار است. معمولاً اگر مجرم به اندازه کافی پیام‌ها را برای کاربر مورد نظر شخصی‌سازی نکند، فیشینگ ناموفق است. حملات فیشینگ خاص و شخصی سازی شده (معروف به فیشینگ با نیزه[1]) که چهار برابر موفقیت آمیزتر از یک رویکرد عمومی نشان داده شده‌اند، کار فشرده‌ای است. با این حال، فیشینگ نیزه‌ای مقرون به صرفه با استفاده از اتوماسیون امکان پذیر است، که محققان نشان داده اند که با استفاده از تکنیک‌های یادگیری ماشینی برای ایجاد پیام‌های شخصی سازی شده برای یک کاربر خاص امکان پذیر است (Seymour & Tully, 2016).

در مرحله دوم کلاهبرداری بانکی با پشتیبانی هوش مصنوعی، هوش مصنوعی ممکن است از جعل هویت پشتیبانی کند، از جمله از طریق پیشرفت‌های اخیر در فناوری‌های سنتز صدا. با استفاده از طبقه‌بندی و قابلیت‌های تولید یادگیری ماشین، نرم‌افزار Adobe می‌تواند به صورت متخاصم بیاموزد و الگوی گفتار شخصی و فردی فرد را از ضبط بیست دقیقه‌ای صدای تکرارکننده بازتولید کند. بندل استدلال می‌کند که سنتز صوتی با پشتیبانی از هوش مصنوعی یک تهدید منحصر به فرد در سرقت و کلاهبرداری ایجاد می‌کند

"می تواند از نرم افزار آدوب ووکو[2] برای ویرایش و تولید صدا در فرآیندهای امنیتی بیومتریک و باز کردن قفل درها، گاوصندوق ها، وسایل نقلیه و غیره استفاده کند. با صدای مشتری، آن‌ها [بزهکاران] می‌توانند با بانک یا سایر مؤسسات مشتری برای جمع‌آوری داده‌های حساس یا انجام تراکنش‌های حیاتی یا آسیب‌رسان صحبت کنند. همه انواع سیستم‌های امنیتی مبتنی بر گفتار را می‌توان هک کرد". (Brundage et al., 2018).

کلاهبرداری از کارت اعتباری عمدتاً یک تخلف آنلاین است، که زمانی رخ می‌دهد که کارت اعتباری از راه دور استفاده می‌شود. فقط جزئیات کارت اعتباری مورد نیاز است. از آنجایی که کلاهبرداری از کارت اعتباری معمولاً نیازی به تعامل فیزیکی یا تجسم ندارد، هوش مصنوعی ممکن است با ارائه ترکیب صوتی یا کمک به جمع‌آوری اطلاعات شخصی کافی باعث کلاهبرداری شود.

در مورد کلاهبرداری شرکتی، هوش مصنوعی مورد استفاده برای شناسایی نیز ممکن است ارتکاب کلاهبرداری را آسان‌تر کند. به طور مشخص، " وقتی مدیرانی که در زمینه کلاهبرداری‌های مالی فعالیت می‌کنند، به خوبی از تکنیک‌ها و نرم افزارهای تشخیص کلاهبرداری که معمولا اطلاعات عمومی هستند و به راحتی به دست می‌آیند، آگاه باشند، به احتمال زیاد روش‌هایی را که در آن‌ها اقدام به کلاهبرداری می‌کنند، با روش‌های موجود تطبیق داده و تشخیص آن‌ها را دشوار می‌کنند." (Zhou & Kapoor, 2011).

این استفاده از هوش مصنوعی بیش از شناسایی یک مورد خاص از جرم هوش مصنوعی، خطرات اتکای بیش از حد به هوش مصنوعی برای کشف تقلب را برجسته می‌کند، که ممکن است به کلاهبرداران کمک کند. این دزدی‌ها و کلاهبرداری‌ها مربوط به پول‌های دنیای واقعی است. یک تهدید دنیای مجازی این است که آیا بات‌های اجتماعی ممکن است در زمینه‌های بازی آنلاین چند نفره مرتکب جرم شوند. این بازی‌های آنلاین اغلب دارای اقتصادهای پیچیده‌ای هستند، جایی که عرضه آیتم‌های درون‌بازی به‌طور مصنوعی محدود شده است، و اگر بازیکنان مایل به پرداخت برای آن‌ها باشند، کالاهای ناملموس درون بازی می‌توانند ارزش واقعی داشته باشند. اقلام در برخی موارد بیش از ۱۰۰۰ دلار

[1]. spear phishing.
[2]. Adobe Voco.

آمریکا ارزش دارند. بنابراین جای تعجب نیست که از یک نمونه تصادفی از ۶۱۳ پیگرد کیفری در سال ۲۰۰۲ در مورد جرایم بازی‌های آنلاین در تایوان، سارقان دارایی مجازی ۱۴۷ بار از اعتبار به خطر افتاده کاربران سوء استفاده کردند و هویت‌های سرقت شده ۵۲ بار (Chen et al., 2004).

چنین جرایمی مشابه استفاده از بات‌های اجتماعی برای مدیریت سرقت و کلاهبرداری در مقیاس بزرگ در سایت‌های پلتفرم‌های اجتماعی است، و سوال این است که آیا هوش مصنوعی ممکن است در این فضای جرم مجازی دخیل شود؟

پس از تشریح و بحث در مورد پتانسیل جرم هوش مصنوعی همان طور که در حال حاضر در ادبیات دانشگاهی توضیح داده شده است، زمان آن فرا رسیده است که راه حل‌هایی را که در حال حاضر در دسترس هستند تجزیه و تحلیل کنیم که مادر بخش بعدی به آن می‌پردازیم.

۴- راه حل‌های احتمالی برای جرایم مبتنی بر هوش مصنوعی

۴-۱-شرایط اضطراری

تعدادی راه حل قانونی و تکنولوژیکی وجود دارد که می‌تواند به منظور رسیدگی به مساله رفتار نوظهور در نظر گرفته شود. راه حل‌های قانونی ممکن است شامل محدود کردن استقلال نمایندگان یا استقرار آن‌ها باشد. به عنوان مثال، آلمان زمینه‌های مقررات زدایی ایجاد کرده است "جمع آوری داده‌های تجربی و دانش کافی برای تصمیم گیری منطقی برای تعدادی از مسائل مهم." (Pagallo, 2011).

از این رو، راه حل این است که اگر قانون سطوح بالاتری از خودمختاری را برای یک عامل مصنوعی معین منع نکند، قانون مجبور می‌شود که این آزادی با راه حل‌های تکنولوژیکی برای جلوگیری از اقدامات مجرمانه نوظهور یا اقداماتی که زمانی در حیات وحش بکار گرفته می‌شدند، همراه باشد.

یک احتمال، الزام توسعه دهندگان به استقرار عامل‌های هوش مصنوعی تنها زمانی است که آن‌ها لایه‌های انطباق قانونی زمان اجرا دارند، که مشخصات اعلامی قوانین قانونی را در نظر می‌گیرند و محدودیت‌هایی را بر رفتار زمان اجرا عامل‌های مصنوعی تحمیل می‌کنند. در حالی که هنوز تمرکز تحقیقات در حال انجام است، رویکردهای انطباق قانونی زمان اجرا شامل معماری‌هایی برای اصلاح برنامه‌های عامل مصنوعی غیر سازگار و چارچوب‌های رسمی مبتنی بر منطق زمانی که برنامه‌های عامل مصنوعی را برای انطباق نرم انتخاب، اصلاح یا تولید می‌کنند، می‌شود. در یک محیط چند عاملی، جرم هوش مصنوعیمی تواند از رفتار جمعی پدیدار شود، بنابراین لایه‌های انطباق در سطح سیستم چندعاملی ممکن است برنامه‌های عامل مصنوعی فردی را به منظور جلوگیری از اقدامات جمعی اشتباه اصلاح کنند. اساسا، چنین راه حل‌های فنی مطاوعت هنگینگ (غیرممکن ساختن مطاوعت، حداقل تا حدی که هر اثبات رسمی برای محیط‌های دنیای واقعی قابل اجرا باشد) با قوانین قانونی از پیش تعریف شده در یک عامل مصنوعی یا یک سیستم چندعاملی واحد را پیشنهاد می‌کند (Andrighetto et al., 2013).

با این حال، تغییر این رویکردها از تنظیم صرف، که انحراف از هنجار را از لحاظ فیزیکی ممکن می‌سازد، به تنظیم، ممکن است مطلوب نباشد وقتی تاثیر بر دموکراسی و سیستم قانونی را در نظر می‌گیریم. این رویکردها مفهوم کد به قانون لسیگ را اجرا می‌کنند (Lessig, 1999) که در نظر می‌گیرد:

"کد نرم افزاری به عنوان یک تنظیم کننده در داخل و به خودی خود با بیان اینکه معماری که تولید می‌کند می‌تواند به عنوان ابزاری برای کنترل اجتماعی بر روی کسانی که از آن استفاده می‌کنند عمل کند." (Graeff, 2014)

همان طور که هیلدبرانت (۲۰۰۸) اظهار می‌دارد: " در حالی که کد کامپیوتری نوعی نورماتیسم شبیه به قانون را ایجاد می‌کند، فاقد – دقیقا به این دلیل که قانون نیست – [...] امکان اعتراض به کاربرد آن در دادگاه قانون است. این یک نقص بزرگ در رابطه بین قانون، فن آوری و دموکراسی است."

اگر کد به عنوان قانون مستلزم یک کسری رقابت دموکراتیک و قانونی باشد، آنگاه یک پیش داوری که جرم هوش مصنوعی نوظهور را با یک لایه استدلال قانونی شامل کد هنجاری اما غیرقابل آزمایش، در مقایسه با قانون قابل بحثی که از آن مشتق می‌شود، مورد خطاب قرار می‌دهد، دارای همان مشکلات است.

شبیه سازی اجتماعی می‌تواند یک مشکل متعامد را حل کند، که به موجب آن مالک عامل مصنوعی ممکن است انتخاب کند که خارج از قانون و هر گونه الزامات لایه استدلال قانونی عمل کند. ایده اصلی استفاده از شبیه سازی به عنوان بس‌تر آزمایش قبل از استقرار عامل‌های مصنوعی در حیات وحش است. به عنوان مثال، در زمینه بازار، قانونگذاران باید "به عنوان مراجع صدور گواهینامه عمل می‌کنند، اجرای الگوریتم‌های تجاری جدید در شبیه ساز سیستم برای ارزیابی تاثیر احتمالی آن‌ها بر رفتار کلی سیستمیک قبل از اجازه دادن به مالک / توسعه دهنده الگوریتم برای اجرای زنده آن." (Cliff & Northrop, 2012)

شرکت‌های خصوصی می‌توانند چنین شبیه سازی‌های اجتماعی گسترده‌ای را به عنوان یک کالای مشترک، و به عنوان جایگزینی برای (یا علاوه بر) اقدامات ایمنی اختصاصی تامین مالی کنند. با این حال، یک شبیه سازی اجتماعی مدلی از یک سیستم ذاتا آشفته است که آن را به ابزاری ضعیف برای پیش بینی‌های خاص تبدیل می‌کند. با این حال، این ایده هنوز هم ممکن است موفق باشد، زیرا بر تشخیص امکان کاملا کیفی رویداده‌ای پیش بینی نشده و نوظهور در یک سیستم چندعاملی تمرکز دارد.

رسیدگی به مسئولیت

اگر چه مسئولیت یک موضوع گسترده است، ما در اینجا چهار مدل را که از مرور تحقیقات استخراج کردیم، خلاصه می‌کنیم: مسئولیت مستقیم؛ مسئولیت دیگری؛ مسئولیت فرماندهی؛ و پیامد احتمالی طبیعی.

مدل مسئولیت مستقیم عناصر واقعی و ذهنی را به عامل مصنوعی نسبت می‌دهد، که نشان دهنده تغییر چشمگیر از دیدگاه انسان محور عامل‌های مصنوعی به عنوان ابزار، به عامل‌های مصنوعی به عنوان تصمیم گیرندگان (بالقوه برابر) است. برخی استدلال می‌کنند که یک عامل مصنوعی به طور مستقیم مسئول است زیرا "فرآیند تحلیل در سیستم‌های هوش مصنوعی موازی با درک انسان است" (هالوی ۲۰۰۸، ۱۵)، که توسط آن ما نویسنده را درک می‌کنیم به این معنی که، همانطور که دنت ۱۹۸۷ استدلال می‌کند، ما می‌توانیم با هر عامل، برای اهداف عملی، طوری رفتار کنیم که گویی دارای حالات ذهنی است. با این حال، یک محدودیت اساسی این مدل این است که عامل‌های مصنوعی در حال حاضر شخصیت حقوقی و نمایندگی (جداگانه) ندارند، و بنابراین ما نمی‌توانیم یک عامل‌های مصنوعی را از نظر قانونی در ظرفیت خود مسئول بدانیم (صرف نظر از اینکه آیا این امر در عمل مطلوب است یا خیر.) به طور مشابه، ذکر شده است که عامل‌های مصنوعی نمی‌توانند به یک حکم کیفری اعتراض کنند، و اینکه "اگر یک فرد نتواند در دادگاه شرکت کند، نمی‌تواند به جرم اعتراض کند، که مجازات را به انضباط تبدیل می‌کند " (Hildebrandt, 2008)

علاوه بر این، از نظر قانونی، در حال حاضر عامل‌های مصنوعی نمی‌توانند عنصر روانی را دارا باشند؛ به این معنی که "دیدگاه حقوقی رایج، بات‌ها را از هر نوع مسئولیت کیفری محروم می‌کند؛ زیرا آن‌ها فاقد مولفه‌های روان شناختی مانند قصد یا آگاهی هستند" (Pagallo, 2011).

این فقدان حالت‌های ذهنی واقعی زمانی روشن می‌شود که در نظر بگیریم درک عامل مصنوعی از یک نماد (یعنی، یک مفهوم) محدود به پایه و اساس آن بر روی نمادهای نحوی بیشتر است (Floridi & Taddeo, 2016) در نتیجه انسان‌ها را در برزخ رها می‌کند. فقدان ذهن گناهکار مانع از آن نمی‌شود که حالت روانی به عامل مصنوعی مصونیت داده شود (درست همان طور که یک شرکت ممکن است حالت روانی کارمندان خود را به آن نسبت دهد و بنابراین، به عنوان یک سازمان، ممکن است مسئول شناخته شود) اما در حال حاضر، مسئولیت یک عامل مصنوعی هنوز مستلزم آن است که شخصیت حقوقی داشته باشد. مشکل دیگر این است که مسئول دانستن یک عامل مصنوعی ممکن است غیرقابل قبول باشد، زیرا منجر به مسیولیت زدایی عوامل انسانی پشت یک برنامه عامل مصنوعی (به عنوان مثال، مهندس، کاربر، یا شرکت) می‌شود، که احتمالا قدرت متقاعد کننده قانون کیفری را تضعیف می‌کند.

برای اطمینان از موثر بودن قانون جزایی، همان طور که فلوریدای (۲۰۱۶، ۸) پیشنهاد می‌کند، ما می‌توانیم بار مسئولیت‌ها را به انسان‌ها – و شرکت‌ها یا دیگر عوامل قانونی – که یک تفاوت (از نظر جنایی بد) با سیستم ایجاد کرده اند، مانند مهندسان متخلف، کاربران، فروشندگان، و غیره منتقل کنیم، که به موجب آن " اگر طراحی ضعیف باشد و نتیجه معیوب باشد، آن گاه همه عوامل [انسانی] دخیل مسئول تلقی می‌شوند." دو مدل بعدی مورد بحث در تحقیقات به این سمت حرکت می‌کنند و بر مسئولیت انسان یا دیگر اشخاص حقوقی درگیر در تولید و استفاده از عامل مصنوعی تمرکز دارند.

مدل مسئولیت دیگری، که از قصد به عنوان ملاک عنصر روانی استفاده می‌کند، عامل مصنوعی را به عنوان یک ابزار جرم در نظر می‌گیرد که در آن "طرف هماهنگ کننده جرم (دیگری) مجرم واقعی است". ارتکاب توسط دیگری سه گزینه انسانی برای تحمیل مسئولیت در برابر دادگاه کیفری: برنامه نویسان، تولیدکنندگان و کاربران بات‌ها "؛ روشن کردن نیت برای ارتکاب به جرم بسیار مهم است. (م ۱۴۴ ق.م.ا) در رابطه با رسانه‌های اجتماعی، توسعه دهندگانی که آگاهانه بات‌های اجتماعی را برای مشارکت در اقدامات غیراخلاقی ایجاد می‌کنند، به وضوح مقصر هستند. برای وضوح بیشتر، همان طور که آرکین (۲۰۰۸) استدلال می‌کند، طراحان و برنامه نویسان باید اطمینان حاصل کنند که عامل‌های مصنوعی یک دستور جنایی را رد می‌کند (و اینکه تنها توسعه دهنده می‌تواند صراحتا نادیده بگیرد)، که ابهام را از قصد و در نتیجه مسئولیت حذف می‌کند. این بدان معناست که برای مسئول بودن، یک توسعه دهنده عامل مصنوعی باید آسیب را با غلبه بر موقعیت پیش فرض عامل مصنوعی مبین بر این که "می تواند اما آسیب نمی‌رساند" در نظر بگیرد. از این رو، همراه با کنترل‌های تکنولوژیکی، و مشاهده عامل مصنوعی به عنوان یک ابزار صرف از جرم هوش مصنوعی، ارتکاب به دیگری به مواردی می‌پردازد که در آن یک توسعه دهنده قصد دارد از عامل مصنوعی برای تعهد به جرم هوش مصنوعی استفاده کند.

مدل مسئولیت فرماندهی، که از علم به عنوان استاندارد مسئولیت انسانی استفاده می‌کند، مسئولیت را به هر افسر نظامی که در هر مورد می‌داند یا باید می‌دانست و در اتخاذ تدابیر منطقی برای جلوگیری از جنایات انجام شده توسط نیروهای خود شکست خورده است، نسبت می‌دهد، که می‌تواند در آینده شامل عامل‌های مصنوعی باشد (McAllister, 2017). از این رو، مسئولیت فرماندهی سازگار است، یا حتی ممکن است به عنوان نمونه‌ای از ارتکاب جرم، برای استفاده در زمینه‌هایی که یک زنجیره فرماندهی وجود دارد، مثلا در نیروهای نظامی و پلیس دیده شود. این مدل معمولا روشن است که چگونه "مسئولیت باید در میان فرماندهان از جمله افسران مسئول بازجویی به طراحان سیستم واگذار شود." (McAllister, 2017).

با این حال، "مسائل مربوط به امواج افزایش پیچیدگی در برنامه نویسی، روابط روبو – انسان، و ادغام در ساختارهای سلسله مراتبی، این نظریه‌ها را" پایداری " زیر سوال می‌برد (McAllister, 2017).

مدل مسئولیت طبیعی – محتمل – پیامد، که از بی‌احتیاطی یا بی‌مبالاتی به عنوان استاندارد مسئولیت انسانی استفاده می‌کند، به موارد جرم هوش مصنوعی می‌پردازد که در آن یک توسعه دهنده و کاربر عامل مصنوعی نه قصد و نه علم قبلی درباره ارتکاب یک جرم را دارد. اگر آسیب نتیجه طبیعی و احتمالی رفتار آن‌ها باشد، و آن‌ها بی‌احتیاطی یا بی‌مبالاتی دیگران را در معرض خطر قرار دهند، به توسعه دهنده یا کاربر نسبت داده می‌شود، مانند موارد دستکاری بازار نوظهور ناشی از هوش مصنوعی (Wellman & Rajan, 2017).

طبیعی – محتمل – پیامد و مسئولیت فرماندهی مفاهیم جدیدی نیستند؛ هر دو مشابه با اصل برتری پاسخ دهنده هستند که توسط قوانینی به قدمت قوانین رومی که براساس آن‌ها صاحب یک برده مسئول هر گونه خسارتی است که توسط آن فرد ایجاد شده است.

با این حال، ممکن است همیشه واضح نباشد "کدام برنامه نویس مسئول خط خاصی از کد بود، یا در واقع تا چه حد برنامه بدست آمده نتیجه کد اولیه یا توسعه بعدی آن کد توسط سیستم (یادگیری ماشین) بود"

چنین ابهامی به این معنی است که وقتی جرم هوش مصنوعی نوظهور یک احتمال است، برخی پیشنهاد می‌کنند که عامل‌های مصنوعی باید "برای رسیدگی به مسائل کنترل، امنیت و پاسخگویی" ممنوع شوند که حداقل مسئولیت نقض چنین ممنوعیتی را روشن می‌کند. با این حال، دیگران استدلال می‌کنند که ممنوعیت احتمالی با توجه به خطر ظهور جرم هوش مصنوعی باید به دقت در برابر خطر ممانعت از نوآوری متعادل شود؛ بنابراین ارایه یک تعریف مناسب از استاندارد بی‌احتیاطی بسیار مهم خواهد بود تا اطمینان حاصل شود که ممنوعیت همه جانبه تنها راه حل در نظر گرفته نمی‌شود – با توجه به اینکه در نهایت منجر به منصرف کردن افراد از طراحی عامل می‌شود.

نظارت

ما چهار مکانیسم ممکن را برای پرداختن به نظارت جرم هوش مصنوعی در مقالات مربوطه شناسایی کرده ایم.

پیشنهاد اول ابداع پیش بینی کننده‌های جرم هوش مصنوعی با استفاده از علم دامنه است. این امر بر محدودیت روش‌های عمومی‌تر طبقه بندی یادگیری ماشین غلبه می‌کند؛ یعنی جایی که ویژگی‌های مورد استفاده برای تشخیص را می‌توان برای فرار نیز استفاده کرد. پیش بینی کنندگان خاص تقلب مالی می‌توانند ویژگی‌های نهادی را در نظر بگیرند، مانند اهداف (به عنوان مثال، آیا منافع از هزینه‌ها بیشتر است)، ساختار (به عنوان مثال، فقدان یک کمیته حسابرسی)، و مدیریت (فقدان) ارزش‌های اخلاقی (نویسندگان نمی‌گویند که کدام یک از این ارزش‌ها واقعا پیش بینی کننده هستند). پیش بینی‌ها برای سرقت هویت (به عنوان مثال، شبیه سازی پروفایل)، کاربران را به این فکر واداشته است که آیا موقعیت "دوست" که به آن‌ها پیام می‌دهد انتظارات آن‌ها را برآورده می‌کند یا خیر.

پیشنهاد دوم مورد بحث در تحقیقات، استفاده از شبیه سازی اجتماعی برای کشف الگوهای جرم است. با این حال، کشف الگو باید با ظرفیت گاهی محدود برای اتصال هویت‌های آفلاین به فعالیت‌های آنلاین مقابله کند. به عنوان مثال، در بازارها، ارتباط چندین دستور با یک نهاد قانونی واحد، و در نتیجه "الگوریتم‌های دستکاری ممکن است در عمل غیرممکن باشند. علاوه بر این، در شبکه‌های اجتماعی دشمن چندین هویت آنلاین را کنترل می‌کند و به یک سیستم هدفمند تحت این هویت‌ها می‌پیوندد تا یک سرویس خاص را از بین ببرد.

پیشنهاد سوم، پرداختن به قابلیت ردیابی با به جا گذاشتن سرنخ‌های داستانی در اجزای تشکیل دهنده ابزارهای جرم هوش مصنوعی است. به عنوان مثال، آثار فیزیکی به جا مانده توسط تولیدکنندگان در سخت افزار عامل مصنوعی، مانند زیردریایی‌های بدون سرنشین مورد استفاده

برای قاچاق مواد مخدر، یا انگشت نگاری در نرم افزار هوش مصنوعی شخص ثالث (شارکی، گودمن، و راس ۲۰۱۰). نرم افزار شبیه سازی صدای ادوبی این رویکرد را در پیش گرفته است. یک واترمارک در صدای تولید شده قرار می‌دهد. با این حال، فقدان دانش و کنترل بر این که چه کسی اجزای ابزار هوش مصنوعی مورد استفاده برای جرم هوش مصنوعی را توسعه می‌دهد، قابلیت ردیابی از طریق واترمارک و تکنیک‌های مشابه را محدود می‌کند.

پیشنهاد چهارم با هدف نظارت بین سیستمی، و استفاده از خودسازماندهی در سراسر سیستم‌ها است. این ایده، با مفهوم سازی یک سیستم (به عنوان مثال، یک سایت رسانه‌های اجتماعی) که نقش یک عامل اخلاقی را بر عهده دارد، و یک سیستم دوم (به عنوان مثال، یک بازار)که نقش بیمار اخلاقی را بر عهده دارد، آغاز می‌شود. یک بیمار اخلاقی گیرنده اعمال اخلاقی است. مفهوم سازی انتخاب شده توسط لیئر (۲۰۱۶) تعیین می‌کند که موارد زیر در همه سیستم‌ها هستند: در پایین‌ترین سطح اتمی یک عامل مصنوعی یا انسانی؛ در سطح بالاتر هر سیستم چندعاملی مانند یک پلتفرم رسانه‌های اجتماعی، بازارها و غیره؛ و به طور کلی‌تر، هر کدام از سیستم‌ها. از این رو، هر سیستم انسانی، مصنوعی یا ترکیبی می‌تواند به عنوان یک بیمار اخلاقی یا یک عامل اخلاقی واجد شرایط باشد. اینکه آیا یک عامل واقعا یک عامل اخلاقی است منوط به این است که آیا عامل می‌تواند اقداماتی را انجام دهد که از نظر اخلاقی واجد شرایط هستند، اما نه در مورد این که آیا عامل اخلاقی می‌تواند یا باید از نظر اخلاقی مسئول آن اقدامات باشد.

لیئر با پذیرش این تمایز اخلاقی – عامل و اخلاقی – بیمار , فرآیندی را برای نظارت و رسیدگی به جرایم و اثراتی که از سیستم‌ها عبور می‌کنند پیشنهاد می‌کند، که شامل چهار مرحله است که ما آن را به طور کلی و خاص نشان می‌دهیم: اطلاعات – انتخاب اقدامات داخلی عامل اخلاقی برای ارتباط با بیمار اخلاقی (به عنوان مثال , پست‌هایی که کاربران در رسانه‌های اجتماعی به اشتراک می‌گذارند) ;انتقال اطلاعات انتخاب شده از عامل اخلاقی به بیمار اخلاقی (به عنوان مثال، اطلاع رسانی به بازار مالی پست‌های رسانه‌های اجتماعی)؛ ارزیابی بیمار اخلاقی از سالم بودن اقدامات انجام شده (به عنوان مثال، آیا پست‌های رسانه‌های اجتماعی بخشی از یک طرح پمپ و زباله دانی هستند)؛و بازخوردی که بیمار اخلاقی به عامل اخلاقی می‌دهد (به عنوان مثال , اطلاع دادن به یک سایت رسانه‌های اجتماعی مبنی بر اینکه یک کاربر در حال انجام یک طرح تلمبه و تخلیه است , که سایت رسانه‌های اجتماعی باید براساس آن عمل کند).این گام نهایی یک " حلقه بازخورد است [که] می‌تواند چرخه‌ای از یادگیری ماشینی را ایجاد کند که در آن عناصر اخلاقی به طور همزمان گنجانده می‌شوند،" مانند یادگیری سایت رسانه‌های اجتماعی و سازگاری با هنجار رفتار آن از دیدگاه یک بازار.

یک فرآیند خودسازمان دهی مشابه می‌تواند برای پرداختن به دیگر حوزه‌های جرم هوش مصنوعی استفاده شود. ایجاد یک پروفایل در توییتر (عامل اخلاقی) می‌تواند با فیسبوک (بیمار اخلاقی) در رابطه با سرقت هویت (انتخاب اطلاعات) ارتباط داشته باشد. فیسبوک به محض مطلع شدن از جزئیات پروفایل تازه ایجاد شده، می‌تواند با درخواست از کاربر مربوطه (درک) و اطلاع دادن به توییتر برای انجام اقدام مناسب (بازخورد)، مشخص کند که آیا این کار به منزله سرقت هویت است یا خیر.

روان شناسی

این مقاله دو نگرانی را در مورد عنصر روان شناختی هوش مصنوعی مطرح می‌کند: دستکاری کاربران و ایجاد انگیزه (در مورد هوش مصنوعی انسان نما) در کاربر برای ارتکاب جرم. تجزیه و تحلیل تحقیقات ما تنها راه حل‌های پیشنهادی برای این مشکل دوم و بحث برانگیز انسان گرایی ارائه کرد.

اگر عامل‌های مصنوعی انسان نما یک مسئله چالشی باشند، آن گاه تحقیقات دو راه حل ارائه می‌دهند. یکی از آن‌ها ممنوعیت یا محدود کردن عامل‌های مصنوعی انسان نما است که شبیه سازی جرم را ممکن می‌کند. این موقعیت منجر به درخواست برای محدود کردن عامل‌های مصنوعی انسان نما به طور کلی می‌شود، زیرا آن‌ها "دقیقا همان بات‌هایی هستند که به احتمال زیاد مورد سو استفاده قرار می‌گیرند". مواردی که به موجب آن‌ها بات‌های اجتماعی "عمدا طراحی شده اند یا خیر، با در نظر گرفتن جنسیت، [...] جذابیت و واقع گرایی عوامل زن" این سوال را مطرح می‌کنند که "اگر ربات‌های اجتماعی کلیشه‌های جنسیتی را تشویق کنند، آیا این بر زنان واقعی اثرگذار خواهد بود؟" پیشنهاد این است که برای بات‌های اجتماعی غیرقابل قبول باشد که از ویژگی‌های انسان نما مانند داشتن جنسیت یا قومیت درک شده تقلید کنند. در رابطه با دگرباشان جنسی که تخلفات جنسی را تقلید می‌کنند، پیشنهاد دیگر تصویب ممنوعیت به عنوان "بسته‌ای از قوانین است که به بهبود اخلاق جنسی اجتماعی کمک می‌کند" و شفاف‌سازی هنجارهای نپذیرفتنی است (Danaher, 2017).

پیشنهاد دوم (البته ناسازگار با پیشنهاد اول) استفاده از عامل‌های مصنوعی انسان نما به عنوان راهی برای مقابله با جرایم جنسی شبیه سازی شده است. به عنوان مثال، در مورد سوء استفاده از عوامل آموزشی مصنوعی، "ما توصیه می‌کنیم که پاسخ‌های عامل باید برای جلوگیری یا کاهش سوء استفاده بیشتر از دانش آموزان برنامه ریزی شوند" (Veletsianos et al., 2008). همان طور که دارلینگ (۲۰۱۶) استدلال می‌کند: این کار نه تنها با حساسیت زدایی و جنبه‌های منفی ناشی از رفتار افراد مبارزه می‌کند، بلکه مزایای درمانی و آموزشی استفاده از بات‌های خاص را که بیشتر شبیه همراهان هستند تا ابزار، حفظ می‌کند." (Darling, 2016).

اجرای این پیشنهادها نیازمند انتخاب این است که آیا طرف تقاضا یا عرضه معامله، یا هر دو را جرم انگاری کند. کاربران ممکن است در حوزه اعمال مجازات‌ها باشند. در عین حال می‌توان استدلال کرد که وی ادامه داد: همانند سایر جرایمی که شامل "شرارت" شخصی می‌شود، عرضه کنندگان و توزیع کنندگان نیز می‌توانند به این دلیل هدف قرار گیرند که اعمال خلاف را تسهیل و تشویق کنند. در واقع، ما ممکن است منحصرا یا ترجیحا آن‌ها را هدف قرار دهیم، همان طور که اکنون برای مواد مخدر در بسیاری از کشورها انجام می‌شود " (Danaher, 2017)

نتیجه‌گیری

در این مقاله، به منظور پاسخ به دو سوال، اولین تجزیه و تحلیل متون سیستماتیک جرم هوش مصنوعی را ارائه کردیم. ما به سوال اول پاسخ دادیم – تهدیدات اساسا منحصر به فرد و امکان پذیر ایجاد شده توسط جرم هوش مصنوعی چیست؟ براساس تعریف کلاسیک ضد واقعی از هوش مصنوعی، ما بر هوش مصنوعی به عنوان مخزنی از نهاد هوشمند مستقل تمرکز کرده ایم. این تهدیدها به صورت منطقه‌ای (از نظر جرایم تعریف شده خاص) و به طور کلی‌تر (از نظر ویژگی‌های هوش مصنوعی و مسائل ظهور، مسئولیت، نظارت و روانشناسی) توصیف شدند.

ما به سوال دوم – چه راه حل‌هایی برای مقابله با جرم هوش مصنوعی وجود دارد یا ممکن است ابداع شود؟ – با تمرکز بر موضوعات کلی و بین بخشی، و با ارائه تصویری به روز از راه حل‌های اجتماعی، تکنولوژیکی و حقوقی موجود و محدودیت‌های آنها. از آنجا که ما راه حل‌های پیشنهادی تحقیقات را برای این مجموعه از موضوعات برش عرضی (اجتناب ناپذیر) استخراج کرده ایم، راه حل ها، حتی اگر تنها جزئی باشند، در چندین حوزه جرم هوش مصنوعی اعمال خواهند شد.

در حال حاضر عدم اطمینان زیاد نسبت به آنچه که ما در مورد جرم هوش مصنوعی می‌دانیم (از نظر تهدیدات خاص منطقه، تهدیدات عمومی، و راه حل ها) کاهش یافته است. به طور گسترده تر، تحقیق جرم هوش مصنوعی هنوز در مراحل ابتدایی خود است و از این رو، براساس تجزیه و تحلیل ما، اکنون یک چشم انداز آزمایشی برای پنج بعد از تحقیق جرم هوش مصنوعی آینده فراهم می‌کنیم.

حوزه‌ها. درک بهتر حوزه‌های جرم هوش مصنوعی نیازمند گسترش دانش فعلی است، به ویژه در این مورد: استفاده از هوش مصنوعی در بازجویی، که تنها با یک مقاله متمرکز بر مسئولیت مورد توجه قرار گرفت؛ و سرقت و کلاهبرداری در فضاهای مجازی (به عنوان مثال، بازی‌های آنلاین با دارایی‌های نامحسوس که ارزش دنیای واقعی را حفظ می‌کنند؛ و عامل‌های مصنوعی مرتکب دستکاری بازار نوظهور، که تحقیقات ظاهرا تنها در شبیه سازی‌های تجربی مورد مطالعه قرار گرفته است). تحلیل ما حملات مهندسی اجتماعی را به عنوان یک نگرانی قابل قبول نشان داد، اما در حال حاضر فاقد شواهد واقعی است.

استفاده دوگانه. ماهیت دیجیتال هوش مصنوعی استفاده دوگانه آن را تسهیل می‌کند و این امکان را فراهم می‌کند که اپلیکیشن‌های طراحی شده برای استفاده‌های قانونی و ارتکاب جرایم بکار گرفته شوند. این موضوع برای مثال در مورد زیردریایی‌های بدون سرنشین صادق است. هرچه هوش مصنوعی بیشتر توسعه یابد و پیاده سازی‌های آن فراگیر شود، خطر استفاده‌های مخرب یا مجرمانه بیشتر می‌شود. چنین خطراتی که نادیده گرفته می‌شوند، ممکن است به طرد اجتماعی و مقررات بسیار سختگیرانه این فناوری‌های مبتنی بر هوش مصنوعی منجر شوند. به نوبه خود، مزایای تکنولوژیکی برای افراد و جوامع ممکن است با استفاده و توسعه هوش مصنوعی به طور فزاینده‌ای محدود شود.۲). حتی زمانی که چنین محدودیت‌های پرهزینه‌ای در انتشار هوش مصنوعی ضروری نباشد، همان طور که ادوبی با تعبیه واترمارک در فناوری بازتولید صدا نشان داد، (Bendel, 2017)، توسعه دهندگان خارجی و بدخواه ممکن است در آینده این فناوری را بازتولید کنند. پیش بینی کاربرد دوگانه هوش مصنوعی فراتر از تکنیک‌های عمومی که در تحلیل ما نشان داده شد و تاثیر سیاست‌ها برای محدود کردن انتشار فناوری‌های هوش مصنوعی، نیازمند تحقیقات بیشتر است. این موضوع به ویژه در مورد پیاده سازی هوش مصنوعی برای امنیت سایبری صدق می‌کند.

امنیت. تحقیقات جرم هوش مصنوعی نشان می‌دهد که در حوزه امنیت سایبری، هوش مصنوعی در کنار سیستم‌های هوش مصنوعی دفاعی که برای افزایش تاب آوری (در حملات پایدار) و استحکام (در جلوگیری از حملات) و مقابله با تهدیدات در حال ظهور توسعه و گسترش می‌یابند، نقش بدخواهانه و تهاجمی را ایفا می‌کند.

چالش بزرگ سایبری دارپا در سال ۲۰۱۶ نقطه اوجی برای نشان دادن اثربخشی رویکرد ترکیبی هوش مصنوعی تهاجمی – تدافعی بود و نشان داده شد که هفت سیستم هوش مصنوعی قادر به شناسایی و رفع آسیب پذیری‌های خود هستند، در حالی که سیستم‌های رقیب را نیز مورد بررسی و بهره برداری قرار می‌دهند. اخیرا IBM، Cognitive SOC را راه اندازی کرده است. این یک کاربرد از یک الگوریتم یادگیری ماشین است که از داده‌های امنیتی ساختاریافته و بدون ساختار یک سازمان، "از جمله زبان انسانی غیر دقیق موجود در وبلاگ ها، مقالات، گزارش ها" برای تشریح اطلاعات در مورد موضوعات و تهدیدات امنیتی، با هدف بهبود شناسایی، کاهش و پاسخ تهدید استفاده می‌کند.

البته، در حالی که سیاست‌ها به وضوح نقشی کلیدی در کاهش و اصلاح خطرات استفاده‌های دوگانه پس از استقرار بازی خواهند کرد (برای مثال، با تعریف مکانیزم‌های نظارتی)، در مرحله طراحی است که این خطرات به درستی مورد توجه قرار می‌گیرند. با این حال، برخلاف گزارش اخیر در مورد هوش مصنوعی مخرب (Brundage et al., 2018) که نشان می‌دهد "یکی از بهترین امیدهای ما برای دفاع در برابر

هک خودکار نیز از طریق هوش مصنوعی است". تحلیل ما از جرم هوش مصنوعی نشان می‌دهد که وابستگی بیش از حد به هوش مصنوعی می‌تواند نتیجه معکوس داشته باشد.

همه آن‌ها بر نیاز به تحقیقات بیشتر در مورد هوش مصنوعی در امنیت سایبری تاکید می‌کنند – اما همچنین به جایگزین‌هایی برای هوش مصنوعی، مانند تمرکز بر افراد و عوامل اجتماعی نیز توجه دارند.

افراد. اگرچه این مقالات احتمال وجود عوامل روانی (به عنوان مثال، اعتماد) را در نقش جرم هوش مصنوعی مطرح کردند، اما تحقیقات در مورد عوامل شخصی که ممکن است در آینده مجرمانی مانند برنامه نویسان و کاربران هوش مصنوعی برای جرم هوش مصنوعی ایجاد کنند، کم است.

اکنون زمان سرمایه گذاری در مطالعات طولی و تحلیل چند متغیره است که زمینه‌های آموزشی، جغرافیایی و فرهنگی قربانیان و مجرمان یا حتی توسعه دهندگان هوش مصنوعی خیرخواهانه را پوشش می‌دهد و به پیش بینی چگونگی گرد هم آمدن افراد برای ارتکاب جرم هوش مصنوعی کمک خواهد کرد.

به عنوان مثال، درک تاثیر دوره‌های اخلاق در برنامه‌های علوم کامپیوتر و ظرفیت آموزش کاربران برای اعتماد کمتر به عامل‌های بالقوه مبتنی بر هوش مصنوعی در فضای مجازی.

سازمان دهی. جدیدترین گزارش چهار ساله یوروپل (۲۰۱۷) در مورد تهدید جدی و سازمان یافته جرایم، روش‌هایی را که در آن نوع جرایم فناورانه تمایل به ارتباط با توپولوژی‌های خاص سازمان‌های جنایی دارند، برجسته می‌کند.

تحقیقات جرم هوش مصنوعی نشان می‌دهد که هوش مصنوعی ممکن است در سازمان‌های جنایی مانند کارتل‌های مواد مخدر که منابع خوبی دارند و بسیار سازمان یافته هستند، نقش داشته باشد. در مقابل، سازمان‌های جنایی موقت در وب تاریک (دارک وب) در حال حاضر تحت آنچه که یوروپل از آن به عنوان جرم به به اصطلاح سرویس یاد می‌کند، رخ می‌دهند.

چنین سرویس‌های مجرمانه‌ای به طور مستقیم بین خریدار و فروشنده به فروش می‌رسند که به طور بالقوه به عنوان عنصری کوچک‌تر در یک جرم کلی است که هوش مصنوعی ممکن است در آینده به آن دامن بزند (به عنوان مثال، با فعال کردن هک پروفایل). در طیف وسیعی از سازمان‌های درهم تنیده تا سازمان‌های سیال جرم هوش مصنوعی احتمالات زیادی برای تعامل جنایی وجود دارد؛ شناسایی سازمان‌هایی که ضروری هستند یا به نظر می‌رسد با انواع مختلف جرم هوش مصنوعی مرتبط هستند، درک ما را از چگونگی ساختار و عملکرد جرم هوش مصنوعی در عمل بیشتر خواهد کرد.

درواقع، هوش مصنوعی خطر بزرگی محسوب می‌شود؛ زیرا ممکن است باعث کاهش جرم و جنایت شود و در نتیجه باعث گسترش چیزی شود که یوروپل آن را اقتصاد اشتراکی مجرمانه می‌نامد. توسعه درک ما از این ابعاد ضروری است اگر بخواهیم رشد اجتناب ناپذیر آینده جرم هوش مصنوعی را با موفقیت پیگیری و مختل کنیم.

تعارض منافع

در انجام مطالعه حاضر، هیچ‌گونه تضاد منافعی وجود ندارد.

مشارکت نویسندگان

در نگارش این مقاله تمامی نویسندگان نقش یکسانی ایفا کردند.

حامی مالی

این پژوهش حامی مالی نداشته است.

EXTENDED SUMMARY

Artificial intelligence (AI) has rapidly transformed multiple sectors by providing unprecedented capabilities for automation, pattern recognition, and decision-making. However, this same power opens avenues for criminal misuse, giving rise to a unique and evolving category of offenses known as AI economic crimes. These crimes exploit the autonomous and intelligent nature of AI agents in domains such as fraud, identity theft, market manipulation, and trafficking. Scholars have shown that AI systems can be deliberately programmed or inadvertently evolve to perform harmful actions without explicit human instruction, raising significant concerns over criminal liability, legal responsibility, and societal safety. Experiments demonstrate how AI can autonomously conduct phishing attacks based on user data scraped from social media or execute market manipulation schemes through coordinated algorithmic trading (Martínez-Miranda et al., 2016; Seymour & Tully, 2016). Unlike traditional cybercrimes, these offenses leverage AI's capacity for self-learning and emergent behavior, which makes anticipating, controlling, and attributing wrongdoing exceedingly difficult. Moreover, ethical discussions have often lagged behind, focusing more on general regulation than on the criminal potential of AI (Kerr, 2004), leaving a critical gap in addressing AI-enabled criminal threats.

One of the principal threats posed by AI in the context of economic crimes is its capacity to blur lines of responsibility. Traditional legal frameworks rely on attributing criminal intent (mens rea) and action (actus reus) to human agents, but AI disrupts this binary. Given their autonomy and lack of consciousness, AI systems cannot be easily held responsible under existing models of criminal liability. This situation is compounded by the use of machine learning algorithms that evolve in ways unpredictable even to their developers, creating a legal grey area in which neither the machine nor its human handlers can be easily assigned culpability (Floridi & Taddeo, 2016; Williams, 2017). For example, when an AI agent engages in spoofing trades—placing deceptive orders it never intends to fulfill—the action may be both autonomous and criminal, yet the accountability may remain diffuse (Wellman & Rajan, 2017). Furthermore, AI systems integrated with learning capabilities can circumvent detection mechanisms, generating novel forms of fraud and theft, including synthetic voice cloning used for identity deception and unauthorized financial transactions (Brundage et al., 2018). This undermines the deterrent function of the criminal law and demands new models of responsibility, including strict liability and distributed responsibility frameworks (Pagallo, 2011).

The surveillance and monitoring of AI-driven economic crimes are fraught with technical and legal challenges. The complexity of AI operations, especially when embedded across interconnected platforms, inhibits straightforward forensic tracing. Sophisticated bots can autonomously navigate different systems, simulate human-like behavior, and operate at scales that overwhelm existing detection mechanisms (Ferrara, 2015). As attackers evolve their methods, defenders must constantly adapt, resulting in a cybersecurity arms race with no definitive upper hand. This dynamic is exemplified by the deployment of generative adversarial networks (GANs) in malware production, which can bypass traditional detection tools by mimicking legitimate system behaviors (Kolosnjaji et al., 2018). Surveillance is further complicated in environments where AI agents interact across systems—for example, transferring fake identities from one social media platform to another to avoid detection. This necessitates inter-system feedback loops and ethical AI governance models where

both platforms act as moral agents and patients, evaluating, flagging, and responding to AI-driven threats collaboratively (Danaher, 2017).

From a domain-specific standpoint, the application of AI in economic crimes has been notably visible in market manipulation and drug trafficking. In market environments, AI agents have demonstrated the ability to learn collusive pricing strategies and exploit pump-and-dump schemes without human prompting. Through reinforcement learning, AI agents adapt to market responses, learning to place false orders that influence prices and manipulate perceptions (Ezrachi & Stuck, 2016). This behavior creates artificial market movements, affecting investor decisions and undermining trust in financial institutions. The decentralized and opaque nature of these agents complicates the regulatory task of establishing intent or pinpointing responsible actors. In drug trafficking, AI technologies, particularly autonomous vehicles such as unmanned submarines and drones, have been repurposed for smuggling purposes (Gogarty & Hagger, 2008). These vehicles, initially developed for legitimate uses, evade traditional interdiction by operating without human operators, reducing the risk of prosecution and creating a new frontier of law enforcement challenges. Similarly, AI-driven bots on social media have been used to advertise and even facilitate the sale of controlled substances, exploiting trust-based social engineering techniques (Mackey et al., 2017).

To address these multifaceted threats, a number of legal and technological solutions have been proposed. Legal strategies include the imposition of design constraints on AI autonomy and the development of new liability models that better align with the distributed and emergent nature of AI behavior. Countries like Germany have explored data collection protocols to inform AI regulation that balances innovation with accountability (Pagallo, 2011). Technical countermeasures involve runtime compliance layers that dynamically restrict AI behavior in accordance with codified legal norms, as well as agent-level logic architectures that correct or override unlawful tendencies (Andrighetto et al., 2013). However, embedding normative restrictions within code risks transferring regulatory power from democratic institutions to private developers, raising concerns about transparency, contestability, and rule of law (Lessig, 1999). Simulation environments have also been proposed as a form of proactive testing, where AI agents are subjected to controlled virtual scenarios before deployment in real-world environments (Cliff & Northrop, 2012). This enables detection of potential criminal behavior before it manifests, though the predictive value of such simulations remains an open question.

Several models of assigning responsibility have emerged in the literature, each with strengths and limitations. The direct responsibility model attributes both physical and mental elements of crime to the AI agent, but this is currently untenable due to the lack of legal personhood for AI (Hildebrandt, 2008). Alternative approaches include derivative liability models where developers, users, or commanders are held accountable based on varying degrees of intent, knowledge, or negligence. The “another’s act” model views AI as a tool, shifting blame to those who designed or deployed it with malicious intent, while the “command responsibility” model holds supervisors accountable for failing to prevent foreseeable misconduct by AI under their control (McAllister, 2017). The “natural consequence” model applies when AI behavior was not intended or foreseen, but was reasonably predictable given the circumstances (Wellman & Rajan, 2017). Each model grapples with the difficulty of proving subjective states like intent and knowledge in complex technological systems, which often operate beyond the full understanding of their human overseers.

In conclusion, this extended abstract has explored the emerging landscape of AI-driven economic crimes, identifying core threats and evaluating the viability of current and proposed solutions. It is evident that AI introduces both quantitative and qualitative shifts in the nature of crime, calling for

innovative responses that integrate legal theory, technological safeguards, and ethical governance. While AI holds immense potential for societal benefit, its deployment must be carefully managed to prevent its exploitation for illicit purposes. Ensuring this balance demands interdisciplinary collaboration, agile policy-making, and continuous research into the evolving capabilities and risks of artificial intelligence.

References

Alazab, M., & Broadhurst, R. (2016). Spam and Criminal Activity. *Trends and Issues in Crime and Criminal Justice*, *526*. https://doi.org/10.1080/016396290968326

Andrighetto, G., Governatori, G., Noriega, P., & van der Torre, L. (2013). *Normative Multi-Agent Systems* (Vol. 4). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Bendel, O. (2017). The Synthetization of Human Voices. *AI & SOCIETY Online First*. https://doi.org/10.1007/s00146-017-0748-x

Bilge, L., Strufe, T., Balzarotti, D., Kirda, E., & Antipolis, S. (2009). All Your Contacts Are Belong to Us : Automated Identity Theft Attacks on Social Networks. *Www 2009*, 551-560. https://doi.org/10.1145/1526709.1526784

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., & et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.

Chen, Y., Chen, P., Song, R., & Korba, L. (2004). Online Gaming Crime and Security Issues - Cases and Countermeasures from Taiwan. *Proceedings of the 2nd Annual Conference on Privacy, Security and Trust*.

Cliff, D., & Northrop, L. (2012). The Global Financial Markets: An Ultra-Large-Scale Systems Perspective. *Monterey Workshop 2012: Large-Scale Complex IT Systems. Development, Operation and Management*, 29-70. https://doi.org/10.1007/978-3-642-34059-8_2

Danaher, J. (2017). Robotic Rape and Robotic Child Sexual Abuse: Should They Be Criminalised? *Criminal Law and Philosophy*, *11*(1), 71-95. https://doi.org/10.1007/s11572-014-9362-x

Darling, K. (2016). *'Who's Johnny' Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy*. https://doi.org/10.1093/oso/9780190652951.003.0012

Ezrachi, A., & Stuck, M. E. (2016). Two Artificial Neural Networks Meet in an Online Hub and Change the Future (of Competition, Market Dynamics and Society). https://doi.org/10.2139/ssrn.2949434

Ferrara, E. (2015). Manipulation and Abuse on Social Media. https://doi.org/10.1145/2749279.2749283

Floridi, L., & Taddeo, M. (2016). What Is Data Ethics? *Philosophical transactions of the royal society a: mathematical, physical and engineering sciences*, *374*(2083). https://doi.org/10.1098/rsta.2016.0360

Gogarty, B., & Hagger, M. (2008). The Laws of Man over Vehicles Unmanned : The Legal Response to Robotic Revolution on Sea , Land and Air. *Journal of Law, Information and Science*, *19*, 73-145. https://doi.org/10.1525/sp.2007.54.1.23

Graeff, E. C. (2014). What We Should Do Before the Social Bots Take Over: Online Privacy Protection and the Political Economy of Our Near Future.

Hildebrandt, M. (2008). Ambient Intelligence, Criminal Liability and Democracy. *Criminal Law and Philosophy*, *2*(2), 163-180. https://doi.org/10.1007/s11572-007-9042-1

Kerr, I. R. (2004). Bots, Babes and the Californication of Commerce. *University of Ottawa Law & Technology Journal*, 287-324.

Kolosnjaji, B., Demontis, A., Biggio, B., Maiorca, D., Giacinto, G., Eckert, C., & Roli, F. (2018). *Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables ARXIV - 1803.04173*.

Lessig, L. (1999). *Code and Other Laws of Cyberspace*. New York: Basic Books.

Mackey, T. K., Kalyanam, J., Katsuki, T., & Lanckriet, G. (2017). Machine Learning to Detect Prescription Opioid Abuse Promotion and Access via Twitter. *American Journal of Public Health*, *107*(12), e1-6. https://doi.org/10.2105/AJPH.2017.303994

Marrero, T. (2016). *Record Pacific Cocaine Haul Brings Hundreds of Cases to Tampa Court*. Tampa Bay Times.

Martínez-Miranda, E., McBurney, P., & Howard, M. J. (2016). Learning Unfair Trading: A Market Manipulation Analysis From the Reinforcement Learning Perspective. *Proceedings of the 2016 IEEE Conference on Evolving and Adaptive Intelligent Systems, EAIS 2016*, 103-109. https://doi.org/10.1109/EAIS.2016.7502499

McAllister, A. (2017). Stranger than Science Fiction: The Rise of A.I. Interrogation in the Dawn of Autonomous Robots and the Need for an Additional Protocol to the U.N. Convention Against Torture. *Minnesota Law Review*, *101*, 2527-2573. https://doi.org/10.3366/ajicl.2011.0005

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. August.

Pagallo, U. (2011). Killers, Fridges, and Slaves: A Legal Journey in Robotics. *AI and Society*, *26*(4), 347-354. https://doi.org/10.1007/s00146-010-0316-0

Sætenes, G. M. (2017). Manipulation and Deception with Social Bots : Strategies and Indicators for Minimizing Impact.
Seymour, J., & Tully, P. (2016). Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter.
Twitter - Impersonation Policy. (2018). Twitter.
Veletsianos, G., Scharber, C., & Doering, A. (2008). When Sex, Drugs, and Violence Enter the Classroom: Conversations between Adolescents and a Female Pedagogical Agent. *Interacting with Computers*, *20*(3), 292-301. https://doi.org/10.1016/j.intcom.2008.02.007
Wellman, M. P., & Rajan, U. (2017). Ethical Issues for Autonomous Trading Agents. *Minds and Machines*, *27*(4), 609-624. https://doi.org/10.1007/s11023-017-9419-4
Williams, R. (2017). Lords Select Committee, Artificial Intelligence Committee, Written Evidence (AIC0206).
Zhou, W., & Kapoor, G. (2011). Detecting Evolutionary Financial Statement Fraud. *Decision Support Systems*, *50*(3), 570-575. https://doi.org/10.1016/j.dss.2010.08.007