Volume 17, Number 4, Winter 2025

Journal of
Science & Technology Policy

Legal Analysis of Algorithmic Deception in the Age of Artificial Intelligence: Challenges and Solutions

Mahdieh Latifzadeh

Assistant Professor of Private Law, The Research Group of Islamic Jurisprudence and Law, Institute for Islamic Studies in Humanities, Ferdowsi University of Mashhad, Mashhad, Iran (latifzadeh@um.ac.ir)

Abstract

The rapid advancement of artificial intelligence (AI) technology in the modern era has created emerging legal issues and challenges. Among the most important of these challenges is the issue of AI deception, and consequently, the production and dissemination of misinformation.This phenomenon is not only a side effect of technological evolution but also indicates a fundamental shift in the dynamics of information dissemination and trust in society. Therefore, addressing this issue is essential because intelligent deception can pose a serious threat to national security, social stability, and civil rights. In light of this necessity, the present study, using a descriptive-analytical approach and comparative study, examines the dimensions of deception in AI. This research, while identifying existing legal challenges, also analyzes appropriate solutions to address this phenomenon. In this regard, the study begins by entering the theoretical foundations and explaining the nature of deception in AI, across four main areas.Then, the mechanism of algorithmic misinformation is structurally analyzed. The next step addresses the legal challenges in dealing with intelligent deception, and finally, legal solutions are presented to counter it. The results of this study reveal that legal systems, especially those with weaker legal infrastructures, face serious challenges and shortcomings in dealing with the emerging complexities of algorithmic deception. These shortcomings mainly stem from the inadequate adaptability of existing laws to rapid technological developments. In such circumstances,effectively addressing the challenges of AI deception requires intelligent utilization of existing legal capacities and a dynamic interpretation of general laws that can be extended. Furthermore, the need to design an innovative and flexible regulatory system, along with the creation of legal frameworks commensurate with the dynamic nature of new technologies, is increasingly felt. This approach can, while addressing immediate needs, provide the necessary platform for the gradual evolution of the Iranian legal system in this area.

Keywords: Technology Regulation, Privacy, Iran's National AI Strategy, Artificial Intelligence

How to Cite this Paper:
Latifzaeh, M. (2025). **Legal Analysis of Algorithmic Deception in the Age of Artificial Intelligence: Challenges and Solutions.** *Journal of Science & Technology Policy*, 17(4), 71-88. {In Persian}.
doi: 10.22034/jstp.2025.11887.1854

Publisher: National Research Institute for Science Policy

سال هفدهم، شماره ٤، زمستان ١٤٠٣

مقاله پژوهشی

فصلنامه علمی-پژوهشی

سیاست علم و فناوری

تحلیل حقوقی فریب الگوریتمی در عصر هوش‌مصنوعی: چالش‌ها و راهکارها

مهدیه لطیف‌زاده

استادیار حقوق خصوصی، گروه پژوهشی فقه و حقوق اسلامی، پژوهشکده مطالعات اسلامی در علوم انسانی دانشگاه فردوسی مشهد، مشهد، ایران.

(latifzadeh@um.ac.ir)

چکیده

پیشرفت سریع فناوری هوش‌مصنوعی در عصر حاضر، چالش‌های حقوقی نوظهوری را ایجاد کرده است. از مهم‌ترین این چالش‌ها، موضوع فریب هوش‌مصنوعی و به‌تبع آن تولید و انتشار اطلاعات نادرست است. این پدیده نه‌تنها یک پیامد جانبی تکامل فناوری محسوب می‌شود بلکه مبین تغییری بنیادین در پویایی انتشار اطلاعات و اعتماد در جامعه است. بدین جهت ضرورت پرداختن به این مسأله ازآن جهت است که فریب هوشمند می‌تواند تهدیدی جدی برای امنیت ملی، ثبات اجتماعی و حقوق شهروندی محسوب شود. به‌دنبال چنین ضرورتی، پژوهش حاضر با رویکرد توصیفی-تحلیلی و مطالعه تطبیقی، به بررسی ابعاد فریب در هوش‌مصنوعی پرداخته است. این پژوهش، ضمن شناسایی چالش‌های حقوقی موجود؛ راهکارهای مناسب برای مواجهه با این پدیده را نیز مورد تحلیل قرار داده است. در این مسیر مطالعه در چهار محور اصلی با ورود به مبانی نظری و تبیین ماهیت فریب در هوش‌مصنوعی آغاز شده و سپس سازوکار گمراه‌سازی الگوریتمی مورد تحلیل ساختاری قرار گرفته است. در گام بعد، چالش‌های حقوقی در مواجهه با فریب هوشمند مورد توجه است و در نهایت نیز، راهکارهای حقوقی برای مقابله ارائه شده است. نتایج حاصل از این پژوهش آشکار می‌سازد که نظام‌های حقوقی - به‌ویژه آن‌هایی که بنیان‌های حقوقی ضعیف‌تری در حوزه هوش‌مصنوعی دارند - در مواجهه با پیچیدگی‌های نوظهور فریب الگوریتمی، با چالش‌ها و کاستی‌های جدی روبه‌رو هستند که عمدتاً از انطباق ناپذیری مناسب قوانین موجود با تحولات سریع فناوری نشأت می‌گیرد. در چنین شرایطی، مقابله مؤثر با چالش‌های فریب هوش‌مصنوعی نیازمند بهره‌برداری هوشمندانه از ظرفیت‌های حقوقی موجود و تفسیر پویا از قوانین عمومیِ قابل تسری، است. افزون بر این، ضرورت طراحی نظام تنظیم‌گری نوآورانه و انعطاف‌پذیر به همراه ایجاد چارچوب‌های قانونی متناسب با ماهیت پویای فناوری‌های نوین، بیش‌ازپیش احساس می‌شود.

کلیدواژه‌ها: تنظیم‌گری فناوری، حریم خصوصی، سند ملی هوش‌مصنوعی، هوش‌مصنوعی.

برای استنادات بعدی به این مقاله، قالب زیر به نویسندگان محترم مقالات پیشنهاد می‌شود:

لطیف‌زاده، مهدیه. (١٤٠٣). **تحلیل حقوقی فریب الگوریتمی در عصر هوش‌مصنوعی: چالش‌ها و راهکارها**. *سیاست علم‌وفناوری*، (٤)١٧، ٨٨-٧١.

doi: 10.22034/jstp.2025.11887.1854

۱- مقدمه

توسعه روزافزون فناوری‌های الگوریتمی به‌ویژه هوش‌مصنوعی، ضرورت توجه به پیامدهای حقوقی و به‌تبع آن، تدوین سیاست‌های تقنینی و تنظیمی در این حوزه را اجتناب‌ناپذیر ساخته است. در این میان، پدیده‌ی «فریب الگوریتمی»[1] به‌عنوان یکی از چالش‌های نوظهور در حوزه تنظیم‌گری فناوری‌های نوین، نیازمند واکاوی دقیق است. فریب الگوریتمی به مجموعه فرآیندهای نظام‌مند در طراحی، پیاده‌سازی یا بهره‌برداری از الگوریتم‌ها اطلاق می‌شود که به هدف دستکاری در پردازش داده‌ها یا سازوکارهای تصمیم‌گیری، منجر به تولید خروجی‌های نادرست یا گمراه‌کننده می‌گردد (تحلیل تفصیلی مؤلفه‌های فنی و حقوقی این مفهوم در بند بعدی ارائه خواهد شد). از سویی دیگر پیچیدگی‌های فنی فناوری‌های الگوریتمی و چالش‌های حقوقی برخاسته از آن، نظام‌های حقوقی را با خلأهای تنظیم‌گری مواجه ساخته است. بر این اساس، مسأله محوری این پژوهش، بررسی پیامدهای حقوقی و ارائه راهکار؛ با توجه به فقدان چارچوب‌های حقوقی و نظام تنظیم‌گری مناسب در مواجهه با پدیده فریب الگوریتمی است. این امر ازآن جهت اهمیت مضاعف می‌یابد که برخلاف فریب در فضای واقعی، فریب الگوریتمی اغلب در لایه‌های پنهان کدنویسی جای گرفته است و تشخیص آن بدون برخورداری از دانش تخصصی دشوار است. علاوه بر این، در الگوریتم‌های یادگیری ماشینی، رفتارهای فریبکارانه ممکن است نه به‌صورت برنامه‌ریزی شده بلکه به‌عنوان پیامدهای ناخواسته در فرآیند یادگیری و به‌مرور زمان بروز یابند. همچنین گستره‌ی عملکرد الگوریتم‌ها در حوزه‌های متعدد ازجمله رسانه‌های اجتماعی، تجارت الکترونیکی، خدمات مالی، نظام سلامت، فرآیندهای مردم‌سالارانه و غیره؛ ابعاد این مسأله را وسیع‌تر ساخته است. درواقع مقیاسِ تأثیرگذاری که می‌تواند به‌صورت همزمان میلیون‌ها کاربر را تحت‌تأثیر قرار دهد، اهمیت تدوین سیاست‌های تنظیم گری کارآمد را دوچندان می‌سازد. در این مسیر، پژوهش حاضر در گام نخست، با تبیین مفهوم فریب در بستر هوش‌مصنوعی و واکاوی مبانی نظری آن، چارچوب مفهومی بحث را ترسیم خواهد نمود. سپس به بررسی سازوکار فریب در هوش‌مصنوعی پرداخته خواهد شد تا درک روشنی از چگونگی وقوع این پدیده حاصل شود. در ادامه، چالش‌های حقوقی ناشی از فریب هوش‌مصنوعی مورد تحلیل قرار می‌گیرد تا در نهایت، با ارائه راهکارهای حقوقی، مسیری برای مواجهه با این چالش‌ها، پیشنهاد شود.

۲- مبانی نظری و چارچوب مفهومی

در عصر کنونی که فناوری با سرعت چشمگیری در حال پیشرفت است، پدیده فریب در سیستم‌های هوش‌مصنوعی به یک چالش جدی تبدیل خواهد شد. برای شناخت این امر، نخست باید به سراغ زیربنای روانشناختی رفتار فریبکارانه در افراد رفت؛ چراکه توجه به این موضوع بینش دقیقی را در مورد چگونگی توسعه و پیاده‌سازی رفتار فریبکارانه در سیستم‌های هوش‌مصنوعی ارائه می‌دهد. توضیح این‌که در رشد شناختی انسان، توانایی فریب دادن یک نقطه عطف مهم محسوب می‌شود که مبین ظهور نظریه ذهن -یعنی ظرفیت درک و پیش‌بینی حالات ذهنی و باورهای دیگران- است [۱]. این ارتباط بین فریب و رشد شناختی به‌ویژه در تحقیقات روانشناسی رشد مشهود است، جایی که مطالعات نشان داده‌اند توانایی کودکان در انجام رفتارهای فریبکارانه موفق معمولاً در حدود چهارسالگی ظاهر می‌شود که نشان‌دهنده درک پیچیده‌ای از باورهای نادرست و دستکاری حالت ذهنی است [۲]. از دیدگاه علوم شناختی، فریب، نیازمند مجموعه‌ای پیچیده از قابلیت‌هاست که فراتر از دستکاری ساده حقیقت است؛ این امر مستلزم درک دیدگاه‌های دیگران، توانایی پیش‌بینی واکنش‌ها و ظرفیت تغییر راهبردی رفتار برای دستیابی به نتایج خاص است [۱]. این موضوع در تحقیقاتِ بسیاری نشان داده شده که کودکان کم‌سن‌تر نیز ممکن است اعمال فریبکارانه انجام دهند لیکن اغلب درکی از تأثیر این اعمال بر باورها و رفتارهای دیگران ندارند [۳]. چارچوب

[1] Algorithmic Deception

نظری درک فریب در حوزه شناختی را می‌توان به‌عنوان انحرافی هدفمند از نمایش صداقت در تعاملات در نظر گرفت. این فرآیند شامل دستکاری آگاهانه اطلاعات برای ایجاد حالات ذهنی خاص در دیگران است. برای فریب مؤثر، درک عمیق از وضعیت ذهنی فعلی هدف و توانایی پیش‌بینی تأثیر اطلاعات دستکاری‌شده ضروری است. فریب علاوه بر نقش دستکاری‌کننده خود، کارکردهای اجتماعی مهمی نیز مانند حفظ هماهنگی در روابط اجتماعی و مدیریت تعاملات پیچیده بین‌فردی دار د. این پدیده همچنین می‌تواند به‌عنوان سازو کاری سازگار کننده عمل کند که در برخی موقعیت‌ها به حفظ روابط اجتماعی و جلوگیری از تعارضات غیرضروری کمک می‌کند. درک این پیچیدگی‌ها و کارکردهای چندگانه فریب برای شناخت بهتر رفتار انسان و تعاملات اجتماعی ضروری است. به زبان ساده فریب، از دیدگاه علمی یک فرآیند شناختی پیچیده است که در آن فرد آگاهانه از حقیقت فاصله می‌گیرد. این فرآیند نیازمند توانایی درک ذهن دیگران و پیش‌بینی واکنش‌های آن‌هاست. برای فریب موفق، شخص باید بتواند حالت ذهنی طرف مقابل را تشخیص دهد و اطلاعات را طوری تغییر دهد که به هدف مورد نظرش برسد. مطالعات نشان می‌دهد فریب تنها یک رفتار منفی نیست بلکه گاهی نقش مهمی در حفظ روابط اجتماعی دارد. برای مثال، برخی دروغ‌های مصلحتی می‌توانند از درگیری‌های غیرضروری جلوگیری کنند. محققین معتقدند این توانایی بخشی از تکامل مغز انسان است که به او کمک می‌کند در موقعیت‌های اجتماعی پیچیده بهتر عمل کند [۱].

فارغ از آنچه در خصوص انسان بیان شد، امروزه با پیشرفت فناوری، این رفتار پیچیده به حوزه‌ای فراتر از تعاملات انسانی گسترش یافته است. سیستم‌های هوش‌مصنوعی با برنامه‌ریزی‌های از پیش تعیین شده و الگوریتم‌های پیشرفته، توانایی‌های جدیدی در زمینه پردازش و ارائه اطلاعات کسب کرده‌اند که می‌تواند منجر به شکل‌گیری نوع جدیدی از فریب شود. توضیح اینکه فریب در هوش‌مصنوعی، مبین تکاملِ پیچیده‌ای از رفتار فریبکارانه انسان است. تعریف سنتی روانشناختی از فریب به‌عنوان ترویج باورهای نادرست، در علوم رایانه‌ای نیز به‌طور گریزناپذیری توسعه یافته است؛ لیکن فریب در هوش‌مصنوعی مفهومی پیچیده‌تر از تعریف سنتی آن در رفتار انسانی است. در خصوص انسان، فریب صرفاً به‌معنای ایجاد باور نادرست در ذهن دیگران است، اما با ظهور هوش‌مصنوعی، این مفهوم تکامل یافته و ابعاد جدیدی پیدا کرده است. [٤].

در دنیای هوش‌مصنوعی، فریب می‌تواند کارکردهای مثبت و سازنده‌ای نیز داشته باشد. برای مثال، یک سیستم هوش‌مصنوعی ممکن است با ارائه اطلاعات گمراه‌کننده به هکرها، از امنیت داده‌های کاربران محافظت کند، یا در شرایط بحرانی با مدیریت هوشمندانه اطلاعات، از آسیب‌های احتمالی جلوگیری نماید. در عین حال نکته مهم این است که ماشین‌ها برخلاف انسان‌ها، قادر به داشتن «باور» به‌معنای واقعی کلمه نیستند. آنچه در هوش‌مصنوعی به‌عنوان فریب شناخته می‌شود، درواقع نوعی دستکاری هدفمند اطلاعات است که می‌تواند حتی در مواقعی به نفع هر دو طرف (سیستم و کاربر) باشد[۵] . بنابراین، مفهوم فریب در حوزه هوش‌مصنوعی را نباید صرفاً به‌عنوان یک رفتار منفی یا مخرب تلقی کرد. این مفهوم می‌تواند به‌عنوان راهبردی هوشمند و کاربردی نیز در نظر گرفته شود که در موارد مثبت و سازنده، مانند تقویت امنیت سایبری، محافظت از داده‌های حساس، بهینه‌سازی عملکرد الگوریتم‌ها و افزایش کارایی سیستم‌های هوشمند، کاربرد دارد [٦]. در عین حال به نظر می‌رسد، ابعاد منفی فریب و به تبع آن نگرانی‌ها در این خصوص بیشتر است. مثلاً سیستم CICERO شرکت متا توانسته است در بستر دیپلماسی با پنهان کردن اطلاعات از هم‌پیمانانش، آنها را فریب دهد [۷]. این رفتار شبیه فریب‌هایی است که در دنیای واقعی –مانند دروغ‌های سیاسی– وجود دارد [۸].

با گسترش مدل‌های زبانی بزرگ، این مشکل جدی‌تر شده است. تحقیقات نشان می‌دهد که این مدل‌ها می‌توانند با تمرین بیشتر، در فریب‌کاری بهتر شوند. به اساس آنچه برخی پژوهش‌ها می‌رساند، میزان فریب‌کاری سیستم‌های هوشمند جدید تا ٤٠ درصد افزایش یافته‌است [۹]. همچنین تولید

محتوای جعلی برخی هوش‌های مصنوعی مثل شبکه‌های مولد تخاصمی[1] که می‌توانند تصاویر و ویدیوهای جعلی بسازند – به‌طوری‌که تشخیص واقعی یا جعلی بودن آن‌هاسخت است– نیز چالش مهمی از مصادیق فریب‌کاری هوش‌مصنوعی است [۱۰].

از نظر تاریخی، مفهوم فریب مصنوعی[2] در اوایل دهه ۲۰۰۰ شکل گرفت و به‌تدریج از نگرانی‌های ساده درباره فریب‌های رایانه‌ای به بحث‌های عمیق‌تر درباره توانایی هوش‌مصنوعی در رفتارهای فریبکارانه تکامل پیدا کرد. امروزه، درک از فریب مصنوعی بسیار پیچیده‌تر شده است. درواقع هوش‌مصنوعی اشکال جدید و پیچیده‌ای از فریب را معرفی کرده است که در سطوح مختلف شناخت و تصمیم‌گیری عمل می‌کند. در این خصوص باید توجه شود که تمایزی اساسی بین دستکاری[3] و فریب[4] در سیستم‌های هوش‌مصنوعی وجود دارد. دستکاری به‌معنای تغییر در فرآیند و ساختار تصمیم‌گیری است، بدون اینکه لزوماً اطلاعات نادرستی ارائه شود. هدف آن تأثیرگذاری بر نحوه تصمیم‌گیری افراد است و از تکنیک‌های روانشناختی و طراحی استفاده می‌کند. برای نمونه استفاده از رنگ‌های خاص برای تحریک احساسات را می‌توان از این موارد دانست. در مقابل، فریب به‌معنای ارائه عمدی اطلاعات نادرست یا گمراه‌کننده است. هدف آن تغییر باورها و تصمیمات افراد با استفاده از داده‌های نادرست است. به‌عنوان نمونه نمایش نظرات جعلی در مورد محصولات یا پنهان کردن هزینه‌های اضافی تا مرحله آخر خرید، از این موارد است. تفاوت اصلی این دو در این است که دستکاری روی «چگونگی» تصمیم‌گیری تأثیر می‌گذارد، در حالی که فریب روی «محتوی و اطلاعاتی» که برای تصمیم‌گیری استفاده می‌شود، تأثیر می‌گذارد. دستکاری معمولاً قابل مشاهده است [هرچند ممکن است ناخودآگاه باشد] اما فریب معمولاً پنهان و غیرقابل تشخیص است. از نظر حقوقی نیز، دستکاری در برخی موارد [مثل بازاریابی] پذیرفته شده است، لیکن فریب معمولاً غیرقانونی و غیراخلاقی محسوب می‌شود. در حوزه هوش‌مصنوعی نیز، گاهی هر دو تکنیک همزمان استفاده می‌شوند مثلاً یک سیستم توصیه‌گر ممکن است هم از تکنیک‌های دستکاری روانشناختی استفاده کند و هم اطلاعات نادرست ارائه دهد [۱۱]. در این میان، دیپ‌فیک‌ها[5] متشکل از دستکاری و فریب محسوب می‌شوند؛ زیرا هم از تکنیک‌های روانشناختی برای تأثیرگذاری بر مخاطب استفاده می‌کنند و هم با ارائه محتوای جعلی، اطلاعات نادرست را به‌شکل‌ی باورپذیر منتقل می‌نمایند. در واقع دیپ‌فیک‌ها نمونه‌ای پیشرفته از فناوری‌های متقاعدکننده[6] هستند که با بهره‌گیری از شبکه‌های مولد تخاصمی و الگوریتم‌های پیچیده هوش‌مصنوعی، قادر به تولید محتوای دیجیتالی غیرواقعی می‌باشند. این سیستم‌ها با پردازش و تحلیل داده‌های بصری و صوتی، می‌توانند تصاویر، ویدئوها و محتوای صوتی را به‌گونه‌ای دستکاری کنند که تشخیصِ اصالت آن‌ها از نمونه‌های واقعی بسیار دشوار است. این فناوری، ظرفیت ایجاد سناریوهای ساختگی با درجه بالایی از واقع‌نمایی را دارد [۱۲-۱۴].

۳- سازوکار فریب در هوش‌مصنوعی

از نظر فنی، فریب در هوش‌مصنوعی می‌تواند به دو روش اصلی انجام شود که شامل شبیه‌سازی و پنهان‌سازی است. در روش شبیه‌سازی، اطلاعات نادرست به‌گونه‌ای ارائه می‌شود که واقعی به نظر برسد، در حالی که در روش پنهان‌سازی، حقیقت از دید مخاطب مخفی نگه‌داشته می‌شود. در اکثر موارد، رفتار فریبکارانه هوش‌مصنوعی مرکب از هر دو روش است تا هدف مورد نظر حاصل شود. توجه به این چارچوب دوگانه به درک اینکه چگونه سیستم‌های هوش‌مصنوعی می‌توانند رفتارهای فریبکارانه را انجام دهند؛ همچنین، درک پیامدهای گسترده این رفتار در تعاملات بین انسان و هوش‌مصنوعی, کمک می‌نماید [۱۵]. پیچیدگی فریب در

[1] Generative Adversarial Networks (GANs)
2 Artificial Deception
[3] Manipulation
[4] Deception
[5] Deepfakes
[6] Persuasive Technologies

هوش‌مصنوعی زمانی بیشتر می‌شود که سیستم‌ها با ترکیب رفتارهای ازپیش تعیین‌شده و توانایی‌های جدید، به دستکاری هدفمند اطلاعات می‌پردازند. بدین ترتیب سازوکار فریب در هوش‌مصنوعی شامل ارائه اطلاعات گمراه‌کننده[1]، خودداری راهبردی از ارائه داده‌ها[2] و شبیه‌سازی رفتارهای انسان‌گونه[3] است [۵]. چنین سازوکاری ریشه در مبانی نظری متعددی دارد که از نظریه بازی[4]، علوم شناختی[5] و روانشناسی تکاملی[6] نشأت می‌گیرد. به مساعدت این مبانی می‌توان درک نمود که چگونه و چرا رفتارهای فریبنده ممکن است در سیستم‌های مصنوعی پدیدار شوند. بنابراین مدل‌سازی نظری رفتار فریبکارانه هوش‌مصنوعی حول چند چارچوب کلیدی شکل گرفته است که جنبه‌های مختلف تعاملات فریبکارانه را پوشش می‌دهند. در این میان، رویکرد نظریه بازی[7] ، به‌ویژه مدل‌های ابربازی[8]، کمک می‌کند تا بتوان درک نمود که چطور یک سیستم هوش مصنوعی می‌تواند از اطلاعات نابرابر برای فریب دادن استفاده کند. البته این مدل در صورتی که مسأله خیلی بزرگ شود، با معضلاتی روبرو می‌شود [۱۶]. برای مدل‌سازی رفتار فریبکارانه در محیط‌های تصادفی — که قطعی نیستند- فرآیندهای تصمیم‌گیری مارکوف[9] به‌عنوان یک چارچوب ظهور کرده‌اند. این مدل عامل‌ها از راهکارهای فریبکارانه یا از طریق اغراق در رفتار به سمت اهداف فریبنده یا با ایجاد ابهام درباره اهداف واقعی خود استفاده می‌کنند [۱۷]. دو دیدگاه نظری مکمل نیز، جنبه‌های شناختی فریب هوش‌مصنوعی را شکل می‌دهد. نظریه دستکاری اطلاعات[10] که بر نحوه دستکاری اطلاعات تمرکز دارد و نظریه فریب بین‌فردی[11] که عوامل اجتماعی-شناختی موثر بر موفقیت فریب را بررسی می‌کند - یعنی توضیح می‌دهد چه عواملی باعث موفقیت در فریب می‌شوند [۹]. با این حال -تلاش برای درک و مدل‌سازی فریب در هوش‌مصنوعی- محققین با یک چالش اساسی روبرو هستند. درواقع با وجود اینکه تعاریف مختلفی در حوزه نظریه بازی و هوش‌مصنوعی نمادین[12] وجود دارد، این نظریه‌ها به تنهایی نمی‌توانند پدیده فریب را در سیستم‌های هوش‌مصنوعی که قابلیت یادگیری دارند، به‌طور کامل توضیح دهند [۱۸،۱۰]. بدین ترتیب پژوهشگران اخیراً رویکرد جدیدی را در پیش گرفته‌اند که به جای تمرکز بر حالات درونی و ذهنی سیستم‌های هوش‌مصنوعی، بیشتر به رفتارهای قابل مشاهده آنها توجه می‌کند. در این رویکرد، مفاهیم قصد[13] و باور [14]، به‌صورت کارکردی تعریف می‌شوند، یعنی بر اساس آنچه که می‌توان در عمل مشاهده و اندازه‌گیری کرد. این تغییرِ رویکرد به سمت تعاریف عملی‌تر و قابل سنجش، گامی مهم در جهت ایجاد یک چارچوب جامع‌تر برای درک فریب در هوش مصنوعی است. با این حال، این حوزه همچنان در حال تکامل است و نیاز به پژوهش‌های بیشتری دارد تا بتواند تمام جنبه‌های پیچیده ی نظری فریب در سیستم‌های هوشمند را پوشش دهد [۱۰].

همچنین باید توجه نمود که توانایی فریب در سیستم‌های هوش مصنوعیِ امروزی، بسیار پیشرفته‌تر از گذشته است. این سیستم‌ها دیگر صرفاً از دستورات ساده پیروی نمی‌کنند، بلکه می‌توانند به‌طور هوشمندانه برنامه‌ریزی و استدلال کنند [۱۹،۱۸]. درواقع، در حال حاضر شاهد تحول از ربات‌های ساده به سیستم‌های خودمختاری می‌باشیم که می‌توانند به‌طور مستقل تصمیم بگیرند و رفتار دیگران را درک کنند [۱۸]. این سیستم‌ها می‌توانند برنامه‌های پیچیده‌ای را در طول زمان اجرا کنند [۲۰]. نکته جالب این است که این سیستم‌ها می‌توانند افکار و واکنش‌های احتمالی دیگران را پیش‌بینی کنند. این قابلیت به خصوص در محیط‌های رقابتی باعث می‌شود که هم سیستم‌های فریب‌دهنده و هم سیستم‌های تشخیص‌دهنده فریب، به‌طور مداوم پیشرفته‌تر شوند [۲۱]. اما یک نکته متناقض وجود دارد؛ اینکه هر چه این سیستم‌ها هوشمندتر

[1] Misleading Information
[2] Strategic Information Withholding
[3] Human-like Behavior Simulation
[4] Game Theory
[5] Cognitive Science
[6] Evolutionary Psychology
[7] Game-theoretic approaches
[8] Hypergame models
[9] Markov Decision Processes (MDPs)
[10] Information Manipulation Theory
[11] Interpersonal Deception Theory
[12] Symbolic AI
[13] Intent
[14] Belief

می‌شوند، در برخی موارد آسیب‌پذیرتر هم می‌شوند. این امر بدین دلیل است که وقتی یک سیستم برای انجام کار خاصی بهینه می‌شود، ممکن است نقاط ضعفی پیدا کند که دیگران بتوانند از آن سوءاستفاده کنند [۲۲]. جریان فریب در سیستم‌های هوش‌مصنوعی از دو مسیر اصلی صورت می‌پذیرد. این موارد برنامه‌نویسی صریح[1] و ویژگی‌های نوظهور حاصل از یادگیری تقویتی [2] است. در برنامه‌نویسی صریح، رفتارهای فریبکارانه مستقیماً در کد برنامه‌نویسی می‌شوند، در حالی که در یادگیری تقویتی، سیستم به طور خودکار این رفتارها را به‌عنوان راه‌حل‌های بهینه کشف می‌کند. با پیشرفت فناوری‌های پردازش زبان طبیعی، سیستم‌ها قادر به یادگیری الگوهای ارتباطی پیچیده‌تر و تقلید رفتارهای انسانی شده‌اند که امکان فریب موفق‌تر را فراهم می‌کند [۲۰]. در نهایت باید دانست که پیاده‌سازی رفتارهای فریبنده در هوش مصنوعی شامل فرآیندی پیچیده و چندمرحله‌ای است که از الگوی سه مرحله‌ای پیروی می‌کند. این امر شامل دور شدن از هدف اصلی، گمراه کردن طرف مقابل، و بازگشت به سمت هدف اصلی است [۲۳].

٤- چالش‌های حقوقی در مواجهه با فریبِ الگوریتمی

فریب در هوض مصنوعی، پرسش‌های بنیادینی را پیش روی نظام‌های حقوقی قرار داده است. در این بند به بررسی دو چالش اساسی در این حوزه پرداخته خواهد شد. نخست، چالش‌های مربوط به احراز قصد فریب در الگوریتم‌ها است که با توجه به ماهیت غیرانسانی و فقدان عناصر متعارف قصد و آگاهی در سیستم‌های هوش‌مصنوعی، چالش‌های قابل توجهی وجود دارد. مورد دوم، چالش‌های برخاسته از ابهامات مربوط به شخصیت هوش‌مصنوعی است که مسأله شناسایی مسئول در اعمال فریبکارانه را با پیچیدگی‌های خاصی مواجه می‌سازد. در نهایت نیز نگاهی مختصر به سایر چالش‌های حقوقی این مسأله، خواهد شد.

٤-١ چالش‌های مربوط به احرازِ قصد فریب در الگوریتم‌ها: پیچیدگی‌های تشخیص عنصر قصد در هوش‌مصنوعی

در بستر حقوق، مفهوم فریب بر پایه رفتار انسانی استوار است که در آن عنصر قصد و آگاهی نقش محوری ایفا می‌کند. قوانین موجود مستلزم احراز قصدِ فریب و وجود آگاهی مرتکب از عمل خود است. در این خصوص به موجب نظام حقوقی ایران، فریب و کلاهبرداری به‌عنوان رفتارهای عمدیِ مبتنی بر قصد انسانی تعریف می‌شوند که تحلیل آنها نیازمند تفکیک دقیق میان مسئولیت کیفری و مسئولیت مدنی است. در حوزه کیفری، تحقق جرایمی مانند کلاهبرداری مستلزم احراز دو بُعد از «عنصر روانی» است. نخست سوءنیت عام است که شامل آگاهی و قصد مرتکب نسبت به انجام عمل مجرمانه می‌شود (مانند ماده ١٤٤ قانون مجازات اسلامی[3] مصوب ۱۳۹۲ – با اصلاحات ١٤٠٣–) و دوم سوءنیت خاص است که به قصدِ رسیدن به نتیجه‌ی خاص اشاره دارد (مانند ماده ۱ قانون تشدید مجازات مرتکبین ارتشاء، اختلاس و کلاهبرداری مصوب ١٣٦٤ – با اصلاحات ١٤٠٣–). علاوه بر این، چنین الزاماتی در ماده ۲۱۷ قانون مجازات اسلامی[4] نیز تأکید شده است و این ماده آگاهی مرتکب از حرمت شرعیِ رفتار را ضروری می‌داند. بنابراین، در حقوق کیفری، قصدِ فریب و آگاهی از ناروا بودن عمل؛ سنگ بنای مسئولیت پذیری است. در مقابل، مسئولیت مدنی ایران در مواجهه با فریب، از دو منظر متمایز قابل بررسی است. از یک سو، مواردی مانند تدلیس در معاملات (ماده ٤٣٨ قانون مدنی مصوب ۱۳۰۷) که به‌عنوان عملی عمدی تعریف می‌شود؛ مستلزم قصدِ فعال برای فریب طرف مقابل مانند پنهان سازی عیوب کالا است. از سوی دیگر، در حوزه مسئولیت مدنی عام (مانند ماده ۱ قانون مسئولیت مدنی ۱۳۳۹) با محوریت مفهوم «تقصیر» حتی در صورت عدم وجود قصد فریب نیز قابلیت جریان دارد. برای نمونه، اگر یک پلتفرم دیجیتال به‌صورت

[1] Explicit Programming
[2] Reinforcement Learning

[3] «در تحقق جرائم عمدی، علاوه بر علم مرتکب به موضوع جرم، احراز قصد او در ارتکاب رفتار مجرمانه ضروری است».

[4] « ... مرتکب باید علاوه بر داشتن علم و قصد، به حرمت شرعی رفتار ارتکابی نیز آگاه باشد».

غیرعمدی اطلاعات نادرست منتشر نماید و به کاربران ضرری وارد شود، مسئولیت جبران خسارت بدون نیاز به اثبات قصد فریب، تنها از طریق احراز بی احتیاطی قابل استناد است. این تفکیک نشان می‌دهد که در حقوق مدنی، برخلاف حقوق کیفری، نیاز به احراز قصدِ فریب همیشگی نیست.

در عین حال گرچه تقصیر و بی‌احتیاطی در نظام حقوقی ایران در حالت متعارف – در بسیاری از موارد– به راحتی قابل احراز است، اما در مورد سیستم‌های هوش‌مصنوعی، تشخیص این موارد نیز پیچیدگی‌های خاص خود را دارد. زیرا حتی با وجود نظارت انسانی، ماهیت پیچیده الگوریتم‌ها و فرآیندهای یادگیری ماشینی؛ تعیین مرز دقیق بین بی‌احتیاطی و خطای سیستمی را دشوار می‌سازد. بدین ترتیب و به طریق اولی، سیستم‌های هوش‌مصنوعی موجود که عمدتاً از نوع هوش‌مصنوعی محدود یا ضعیف[1] هستند، علی‌رغم اینکه تحت نظارت انسانی فعالیت می‌کنند، فاقد قصد و آگاهی به‌معنای متعارف انسانی هستند و صرفاً بر اساس الگوریتم‌ها و داده‌های آموزشی عمل می‌کنند. حتی در صورت توسعه یعنی استقرار هوش‌مصنوعی قوی[2] –که توانایی درک و تصمیم‌گیری نزدیک به انسان را خواهد داشت– باز هم مسأله قصد و آگاهی به شکل انسانیِ آن، محل بحث خواهد بود. این موضوع، چه در حالت فعلی (وجود هوش‌مصنوعی ضعیف) و چه در حالت پیشرفته‌تر (تحقق هوش‌مصنوعی قوی) چالش‌های خاص خود را دارد. درواقع توسعه هوش‌مصنوعی، مسائل پیچیده‌ای را در زمینه احراز قصد فریب و به تبع آن تعیین مسئول و انتساب عمل به اشخاص، ایجاد می‌کند. بدین جهت نظام‌های حقوقی متعارف که بر مبنای درک متقابل از ذهن و باورهای طرفین بنا شده‌اند، در مواجهه با هوش‌مصنوعی با چالش مواجه می‌شوند. این امر بدین دلیل است که نمی‌توان چنین درکی را به سیستم‌های هوش‌مصنوعی نسبت داد. این وضعیت نیازمند بازنگری در چارچوب‌های حقوقی موجود و ایجاد رویکردهای جدید برای تعریف و اثبات فریب توسط هوش‌مصنوعی باشد که کمتر بر مفاهیم سنتی مربوط به قصد و آگاهی، متکی باشد [۲۵،۲٤].

فارغ از آنچه بیان شد همچنین نمی‌توان فریب ناشی از سیستم‌های هوش‌مصنوعی را معادل کلاهبرداری انسانی دانست. این امر بدین دلیل است که در نظام حقوقی ایران، جرم کلاهبرداری، یک جرم عمدیِ خاص است که مستلزم احراز دو رکن سوءنیت عام (قصد انجام عمل فریبکارانه) و سوءنیت خاص (قصد نتیجه) است. حال آنکه هوش‌مصنوعی به مثابه مجموعه‌ای از الگوریتم‌ها، فاقد «اراده‌ی آزاد» و «آگاهی» برای شکل دادن به این سوءنیت است. برای مثال، اگر یک سیستم پیش بینی کننده ی مالی مبتنی بر هوش‌مصنوعی به دلیل خطای برنامه نویسی یا داده‌های ناقص، پیش بینی نادرستی ارائه دهد و سرمایه گذاران را به اشتباه اندازد، هرچند نتیجه‌ی این رفتار شبیه کلاهبرداری است، اما به دلیل فقدان قصد ذاتی سودجویی در الگوریتم، نمی‌توان آن را مصداق جرم کلاهبرداری دانست. در اینجا، مسأله اصلی انتقال مسئولیت از «الگوریتم» به «انسان» است. پس از منظر حقوق کیفری، عدم امکان انتساب مسئولیت به هوش‌مصنوعی یک محدودیت ماهوی است، زیرا تحقق جرائم عمدی منوط به وجود «علم و قصد مرتکب» است. با این حال، حقوق مدنی با تکیه بر مفاهیمی مانند تقصیر و تسبیب، راهی برای جبران خسارت‌های ناشی از هوش‌مصنوعی ارائه می‌نماید. برای نمونه، در جایی که شرکت ارائه دهنده ی یک چت بات مالی، به دلیل غفلت در آزمایش سیستم یا استفاده از داده‌های مغرضانه، موجب فریب کاربران شود، می‌توان مسئولیت مدنی این شرکت را بر اساس بی احتیاطی محقق دانست، حتی اگر هیچ قصد فعالی برای فریب وجود نداشته باشد. نکته‌ی کلیدی که باید به آن تأکید کرد، تمایز میان «فعل فریبنده» و «جرم کلاهبرداری» است. درست است که رفتار یک الگوریتم ممکن است عملاً موجب فریب شود، اما تا زمانی که قصد انسانی پشت این رفتار احراز نشود، نمی‌توان از چارچوب کیفریِ کلاهبرداری استفاده کرد. اینجاست که قانونگذار باید با نگاهی پیشگیرانه، به سمت تدوین قوانین خاص برای مدیریت خطرات ناشی از

1 Artificial Narrow Intelligence / Weak AI

2 Artificial General Intelligence / Strong AI

هوش‌مصنوعی حرکت کند.

ضرورت احراز قصدِ فریب در مورد سیستم‌های هوش‌مصنوعی، موضوع را به سمت چالش عمیق‌تری در حوزه حقوق هوش‌مصنوعی هدایت می‌کند. این مسأله؛ عدم امکان انتساب قصد مجرمانه به الگوریتم‌ها است که ریشه در فقدان شخصیت مستقل حقوقی برای هوش‌مصنوعی دارد. به‌طور کلی مسأله شخصیتِ هوش‌مصنوعی، ارتباط مستقیمی با بحث فریبندگی الگوریتم‌ها دارد؛ زیرا در صورت پذیرش شخصیت برای سیستم‌های هوشمند، می‌توان مسئولیت اعمال فریبنده را مستقیماً متوجه خود سیستم دانست. اما در غیر این فرض، باید به دنبال انتساب مسئولیت به اشخاص حقیقی یا حقوقی دیگری بود. همچنین این چالش حقوقی زمانی پیچیده‌تر می‌شود که الگوریتم‌های هوش‌مصنوعی به‌صورت خودمختار و بدون دخالت مستقیم انسان، رفتارهای فریبنده از خود نشان می‌دهند. در چنین شرایطی، عدم وجود شخصیت حقوقی مستقل برای هوش‌مصنوعی، خلأ قانونی جدی در زمینه تعیین مسئول برای جبران خسارت و اعمال ضمانت اجرا ایجاد می‌کند. چالش‌های مربوط به شخصیت به تفصیل در ادامه بررسی خواهند شد.

٤-٢ چالش‌های مربوط به شخصیت هوش‌مصنوعی: شناسایی مسئول در اَعمال فریبکارانه

ابهام در خصوص وجود یا فقدان شخصیت برای سیستم‌های هوش‌مصنوعی؛ پیچیدگی مسأله را افزون می‌کند [٤٠]. از منظر حقوقی، مسأله شخصیت سیستم‌های هوش‌مصنوعی به‌عنوان یکی از پیچیده‌ترین چالش‌های حقوقی معاصر مطرح است که می‌توان آن را اینگونه تبیین نمود[1]. در نظام‌های حقوقی کنونی، شخصیت به دو دسته اشخاص حقیقی و حقوقی تقسیم می‌شود که هوش‌مصنوعی در هیچ یک از این دو دسته‌بندی نمی‌گنجد. بدین جهت در حالی که هوش‌مصنوعی ممکن است به‌عنوان یک شخص حقوقی اصلی مانند یک شخص حقیقی تلقی نشود، مفهوم «شخص حقوقی مشتق»[2] به‌عنوان چارچوبی بالقوه برای پاسخ‌گویی به چالش‌های مربوط به هوش‌مصنوعی مطرح شده است. مفهوم «شخص حقوقی مشتق» یعنی هوش‌مصنوعی بتواند واجد برخی حقوق و تکالیف محدود باشد، بدون آنکه از تمامی حقوق و تکالیف یک شخص حقوقی کامل برخوردار گردد. این شخصیت حقوقی مشتق، از شخصیت حقوقی سازمان یا شرکت مادر نشأت می‌گیرد و به‌صورت تبعی به سیستم هوش‌مصنوعی تسری می‌یابد. در این چارچوب، مسئولیت‌های حقوقی به‌صورت سلسله مراتبی تقسیم می‌شود؛ یعنی توسعه‌دهنده به‌عنوان شخص حقوقی اصلی، مسئولیت اولیه را بر عهده دارد؛ سیستم هوش‌مصنوعی به‌عنوان شخص حقوقی مشتق، در محدوده مشخصی مسئول شناخته می‌شود و کاربران نهایی نیز در صورت سوء استفاده یا نقض دستورالعمل‌ها مسئول خواهند بود. این ساختار حقوقی در بستر مسئولیت تضامنی می‌تواند مبنایی برای جبران خسارات احتمالی باشد [٢٦]. در مقابل باید توجه نمود که آنچه بیان شد؛ یکی از نظرات موجود در خصوص شخصیت هوش‌مصنوعی است. گرچه در سال‌های ۲۰۱۷ تا حدود سال ۲۰۲۱ (زمان ارائه پیشنهاده قانون هوش‌مصنوعی اتحادیه اروپا) نظریاتی مانند «شخصیت حقوقی مشتق» مطرح شده است، لیکن امروزه قانونگذاران نظام‌های حقوقی پیشتاز در حوزه هوش‌مصنوعی؛ رویکردهای دیگری را برگزیده‌اند. ناگفته نماند، مبدا این نظر قطعنامه پارلمان اروپا با عنوان «قواعد قانون مدنی در حوزه رباتیک[3]» مصوب ۲۰۱۷ است که به موجب آن ایده اعطای «شخصیت الکترونیکی[4]» به ربات‌های پیشرفته مطرح شده بود. با وجود این، در حال حاضر، اتحادیه اروپا در قانون هوش‌مصنوعی[5] مصوب ٢٠٢٤،

[1] مسأله شخصیت هوش مصنوعی و چالش های ناشی از آن، در آثار علمی فارسی نیز مورد توجه قرار گرفته است [۴۰].

[2] Derivative Legal Subject

[3] Civil Law Rules on Robotics
uropean Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))
https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52017IP0051

[4] Electronic Personhood

[5] Artificial Intelligence Act
European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and

به جای اعطای شخصیت حقوقی به سیستم‌های هوش‌مصنوعی، رویکرد «مسئولیت زنجیره‌ای[1]» را پذیرفته است (ماده ۲۵ قانون هوش‌مصنوعی اتحادیه اروپا مشاهده شود). در این رویکرد، مسئولیت‌ها به‌صورت نظام‌مند میان تمامی ذینفعان ازجمله توسعه‌دهندگان، تأمین‌کنندگان و کاربران تقسیم می‌شود. همچنین با اتخاذ رویکرد مبتنی بر خطر[2] سطوح مختلف مسئولیت بر اساس میزان خطرات بالقوه سیستم‌های هوش‌مصنوعی تعیین می‌گردد. توضیح دقیق‌تر اینکه زنجیره مسئولیت شامل توزیع‌کننده[3]، واردکننده[4]، توسعه دهنده (به کار گیرنده)[5] ، شخص ثالث[6] –تمامی این اشخاص به‌عنوان ارائه دهنده[7] تلقی می‌شوند– ارائه دهنده اولیه[8] و تولیدکننده محصول[9] می‌باشد. مسئولیت این اشخاص در جایی است که نام یا علامت تجاری خود را روی سیستم قرار دهند؛ تغییرات اساسی در سیستم ایجاد کنند، یا هدف سیستم را به‌گونه‌ای تغییر دهند که به سیستم پرخطر تبدیل شود. در این ساختار، مسئولیت از نوع نسبی[10]– نه تضامنی[11]– است. این امر بدان معناست که هر شخص به میزان مداخله و نقش خود در زنجیره‌ی ارزش مسئول شناخته می‌شود. همچنین ارائه دهنده اولیه موظف به همکاری با ارائه دهندگان جدید[12] است. تولیدکننده محصول نیز مسئولیت ایمنی محصولات را بر عهده دارد و اشخاص ثالث هم مسئول تأمین ابزارها، خدمات و اجزای مورد نیاز هستند. به علاوه مسئولیت‌ها قابلیت انتقال[13] دارند و با انتقال به ارائه دهنده جدید، مسئولیت ارائه دهنده اولیه ساقط می‌شود[14]. بدین ترتیب این مسئولیت‌ها به‌صورت زنجیره‌ای و مرحله به مرحله تعریف شده‌اند [۲۷].

این رویکرد قانونی اتحادیه اروپا در زمینه مسئولیت زنجیره‌ای، مبین تحول مهمی در نظام‌های حقوقی در برخورد با هوش‌مصنوعی است. با این حال، چالش‌های نوظهور در حوزه هوش‌مصنوعی نیازمند انطباق بیشتر ساختارهای حقوقی موجود است. در این راستا، ساختارهای حقوقی فعلی باید با سازوکارهای پاسخگویی غیرمستقیم[15] تطبیق یابند. این امر یعنی مسئولیت حقوقی در حوزه هوش‌مصنوعی عمدتاً غیرمستقیم باشد و متوجه سازندگان [مسئولیت می‌تواند شامل طراحی نامناسب سیستم یا عدم پیش‌بینی تمهیدات امنیتی کافی باشد]، توسعه‌دهندگان و بهره‌برداران گردد. این مسئولیت در دو بعد مدنی (جبران خسارت) و کیفری (مجازات) قابل اعمال است [۲٤]. بدین ترتیب در خصوص مصادیق فریب در هوش‌مصنوعی مانند کلاهبرداری فیشینگ[16] [نوعی کلاهبرداری رایانه‌ای] نیز باید قوانین موجود تطبیق و توسعه –برای پوشش موارد جدید– یابند [۲۵].

٤–۳ سایر چالش‌های فریب در هوش‌مصنوعی

فارغ از تبیین کلی چالش‌ها، اشاره‌ای مختصر به سایر نگرانی‌های این موضوع– مستند به پرونده‌های حقوقی موجود– نیز خالی از فایده نیست. از مهمترین این موارد افترا و ارائه اطلاعات نادرست[17] است. توضیح اینکه دعاوی متعددی در قالب افترا علیه شرکت OpenAI سازنده ChatGPT به دلیل تولید و ارائه اطلاعات نادرست در مورد اقدامات مجرمانه و مسائل حقوقی افراد مطرح شده است [۲۸] . همچنین مواردی که سیستم‌های هوش‌مصنوعی محتوی نادرست را در مقیاس وسیع تولید نمودند و بر حوزه‌های مختلف حقوقی ازجمله امنیت ملی و حمایت از مصرف‌کننده، موثر بودند [۲۹]. به بیان ساده در سال ۲۰۲۳، مارک دیویس[18] یک استاد دانشگاه، متوجه شد که ChatGPT اطلاعات نادرستی درباره پیشینه حرفه‌ای او تولید کرده و او

(EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union. Retrieved from https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[1] Chain of Responsibility
[2] Risk-based Approach
[3] Distributor
[4] Importer
[5] Deployer
[6] Third-party
[7] provider
[8] Initial Provider
[9] Product Manufacturer
[10] Several Liability
[11] Joint Liability
[12] New Providers
[13] Transferability
[14] Liability Discharge
[15] Indirect Accountability Mechanisms
[16] Phishing
[17] Defamation and False Information Generation
[18] Mark Davies

را به کلاهبرداری متهم کرده است. این نخستین موج از موارد مشابهی بود که توجه جامعه حقوقی را به خود جلب کرد و حقوقدانان با پرونده‌های مشابه بیشتری مواجه شدند. هوش‌مصنوعی OpenAI گاهی داستان‌هایی خیالی درباره افراد می‌ساخت که شامل اتهامات جعلی، سوابق کیفری ساختگی و حتی روابط تجاری غیرواقعی می‌شد. این مسأله به‌ویژه برای افراد شناخته‌شده و صاحبان کسب‌وکار بسیار مشکل‌ساز شد. مطالعات نشان می‌دهد که این مشکل فراتر از موارد فردی است. به موجب گزارش‌ها سیستم‌های هوش‌مصنوعی می‌توانند در عرض چند ساعت هزاران مطلب نادرست تولید کنند که تشخیص درستی یا نادرستی آنها برای کاربران عادی تقریباً غیرممکن است. این وضعیت هنگامی پیچیده‌تر شد که مشخص شد محتوی تولیدشده توسط هوش‌مصنوعی می‌تواند امنیت ملی را نیز تهدید کند. برای مثال، انتشار اخبار جعلی درباره تصمیمات سیاسی یا نظامی، یا اطلاعات نادرست درباره زیرساخت‌های حیاتی کشور از این موارد است. همچنین در حوزه حمایت از مصرف‌کننده نیز، موارد متعددی از تولید نظرات جعلی محصولات، توصیه‌های گمراه‌کننده سرمایه‌گذاری و حتی مشاوره‌های پزشکی نادرست توسط هوش‌مصنوعی گزارش شده است. وجود این مسائل، باید دغدغه نهادهای نظارتی و قانون‌گذار را در خصوص تنظیم گری هوش‌مصنوعی به حداکثر برساند [۳۰]. همچنین دستکاری شواهد در رسیدگی‌های قضایی[1] نیز از مصادیق فریب است که به چالشی مهم تبدیل شده است. توضیح اینکه پیشرفت هوش‌مصنوعی در تولید و دستکاری محتوی دیجیتال، چالش‌های جدی برای نظام‌های قضایی ایجاد کرده است. این فناوری توانایی تغییر و جعل شواهد دیجیتال را به شکلی پیشرفته فراهم می‌کند که تشخیص اصالت آنها را دشوار می‌سازد. در پرونده‌های قضایی که به شهادت شهود و مدارک دیجیتالی متکی هستند، این مسأله اهمیت ویژه‌ای پیدا می‌کند. برای مقابله با این چالش، سیستم‌های قضایی نیازمند پیاده‌سازی روش‌های پیشرفته تشخیص محتوای تولیدشده توسط هوش‌مصنوعی هستند [۳۱].

۵- راهکارهای حقوقی: به سوی چارچوبی برای پاسخ‌گویی

مقابله با چالش‌های فریب توسط هوش‌مصنوعی نیازمند رویکردی دوگانه است که از یک سو بر ظرفیت‌های موجود نظام‌های حقوقی تکیه می‌کند و از سوی دیگر با تنظیم‌گری هدفمند به مواجهه با ماهیت پویای فناوری‌های نوین می‌پردازد. این دو راهکار در ادامه تشریح خواهد شد.

۵-۱ بهره‌مندی از چارچوب‌های حقوقی موجود

با توجه به اینکه کشورهایی همچون ایالات متحده آمریکا و اتحادیه اروپا، نه تنها پیشگام در توسعه فناوری هوش‌مصنوعی هستند، بلکه زیرساخت‌های حقوقی تقریبا مناسب و تجربه طولانی نیز در تنظیم‌گری فناوری‌های نوین دارند؛ بررسی رویکرد این نظام‌های حقوقی اهمیت ویژه‌ای می‌یابد. این در حالی است که بسیاری از کشورها –ازجمله ایران– هنوز در مراحل اولیه شناخت و مواجهه با چالش‌های مربوط به هوش‌مصنوعی و بالتبع تنظیم گری این حوزه هستند. بدین ترتیب ابتدا به رویکرد ایالت متحده آمریکا سپس اتحادیه اروپا و در نهایت به نظام حقوقی ایران اشاره خواهد شد.

ایالات متحده با توجه به پیچیدگی‌های روزافزون فناوری هوش‌مصنوعی، به ویژه در حوزه فریب، علاوه بر قوانین عمومی موجود و مرتبط، اقدام به تدوین قوانین جدید و به‌روزرسانی چارچوب‌های قانونی خود کرده است. توضیح اینکه بیشتر ایالت‌ها، قوانین عمومی مربوط به جعل هویت دارند که پیش از ظهور اینترنت تصویب شده‌اند، لیکن این قوانین به‌طور بالقوه می‌توانند به تمام انواع رسانه‌ها، ازجمله ارتباطات آنلاین و ویدیوها حتی هوش‌مصنوعی اعمال شوند. همچنین قوانین مربوط به جعل هویت آنلاین قایل اشاره است. با رایج شدن شبکه‌های اجتماعی و ایمیل، قانون‌گذاران در حداقل ۱۷ ایالت، قوانینی تصویب کرده‌اند که به‌طور خاص به جعل هویت آنلاین اشاره دارند؛ جعل هویتی که با هدف ارعاب، زورگویی، تهدید یا آزار افراد از طریق شبکه‌های اجتماعی، ایمیل یا دیگر ابزارهای ارتباطاتی

[1] Evidence Manipulation in Legal Proceedings

الکترونیکی یا آنلاین انجام می‌شود. این ایالت‌ها شامل کالیفرنیا، کنتیکت، فلوریدا، هاوایی، ایلینوی، لوئیزیانا، ماساچوست، می‌سی‌سی‌پی، نیوجرسی، نیویورک، کارولینای شمالی، اوکلاهما، رود آیلند، تگزاس، یوتا، واشنگتن و وایومینگ هستند. با وجود این به‌طور خاص، قوانین مربوط به دیپ‌فیک از سال ۲۰۱۹، در چند ایالت، تصویب شده است که استفاده از دیپ‌فیک‌ها را هدف قرار می‌دهند. این قوانین به‌طور انحصاری به دیپ‌فیک‌هایی که توسط هوش‌مصنوعی ایجاد شده‌اند محدود نمی‌شوند، بلکه به‌طور گسترده‌تر به تصاویر صوتی یا تصویری دستکاری‌شده فریبنده اشاره دارند که با نیت بدخواهانه ساخته شده و دیگران را بدون رضایت آن‌ها به‌طور کاذب نمایش می‌دهند. همچنین حداقل ۴۰ ایالت در جلسات قانون‌گذاری سال ۲۰۲٤ قوانین مربوط به دیپ‌فیک را در دست بررسی دارند. تاکنون حداقل ۵۰ لایحه تصویب شده است [۳۲] .

بدین ترتیب در ایالات متحده، قانون فدرال جامعی که به‌طور خاص به دیپ‌فیک‌ها بپردازد، وجود ندارد و ایالت‌ها با ابتکار عمل خود برای قانون گذاری اقدام نموده‌اند. به‌عنوان مثال «قانون تگزاس مربوط به مقابله با دیپ فیک ها در انتخابات مصوب ۲۰۱۹» [1] به‌طور خاص به موضوع دیپ‌فیک‌ها در زمینه انتخابات می‌پردازد که از اولین تلاش‌های حقوقی در سطح ایالتی آمریکا برای مقابله با تهدید دیپ‌فیک‌ها در روند انتخابات بود. هدف اصلی آن جرم‌انگاری ساخت و انتشار ویدیوهای دستکاری شده با هدف تأثیرگذاری بر نتایج انتخابات است. همچنین قانون رسانه‌های صوتی یا تصویری فریبنده کالیفرنیا[2] نیز مشابه قانون تگزاس است و به موضوع فریب رسانه‌ها در انتخابات می‌پردازد، اما جزئیات و محدوده آن متفاوت است. قانون کالیفرنیا به‌طور خاص به رسانه‌های صوتی و تصویری گمراه‌کننده در زمینه انتخابات می‌پردازد [۳۳]. به علاوه قانون دفاع از دموکراسی در برابر فریب دیپ‌فیک مصوب[3] ۲۰۲٤ ایالت کالیفرنیا نیز وجود دارد. این قانون با هدف محافظت از یکپارچگی دموکراسی و جلوگیری از انتشار رسانه‌های دستکاری‌شده و اطلاعات نادرست در فضای آنلاین که می‌تواند رأی‌دهندگان را فریب داده یا بر فرآیند رأی‌گیری تأثیر بگذارد، به تصویب رسیده است[4]. لازم به ذکر است، این اسناد فارغ از چارچوب‌های حقوقی است که به‌طور کلی در خصوص تنظیم گری هوش‌مصنوعی[5] قابل استفاده‌اند.

با توجه به آنچه بیان شد، اقدامات قانونی ایالات متحده در مقابله با فریب هوشمند، به سه بُعد مختلف تقسیم می‌شود. بُعد نخست شامل قوانین عمومی است که برای جعل هویت و... وضع شده‌اند و قابلیت اعمال به فضای آنلاین و هوش‌مصنوعی را نیز دارند. بُعد دوم مربوط به قوانین خاص دیپ‌فیک است که از سال ۲۰۱۹ در برخی ایالت‌ها مانند تگزاس و کالیفرنیا تصویب شده و به‌طور ویژه به محتوای دستکاری شده توسط هوش‌مصنوعی می‌پردازند. بُعد سوم نیز شامل چارچوب‌های حقوقی کلی مرتبط با هوش‌مصنوعی است که به‌صورت عام به موضوع هوش‌مصنوعی می‌پردازد و از دو دسته قبلی متمایز است. بدین جهت ایالت‌ها در مقابله با فریب هوشمند پیشگام بوده‌اند، در حالی که سطح فدرال عمدتاً به قوانین عمومی موجود اکتفا کرده است. این برتری عملکرد ایالت‌ها نیز به دلیل انعطاف‌پذیری بیشتر، نزدیکی به

---

1 Texas Senate Bill 751 (SB 751): Relating to the Regulation of Deepfakes in Elections. (2019).
URL: https://legiscan.com/TX/text/SB751/id/1902830

2 California Legislative Information. (2019). “AB-730 Elections: deceptive audio or visual media.”
URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

3 Defending Democracy from Deepfake Deception Act of 2024
https://legiscan.com/CA/text/AB2655/id/2987162

4 برای دسترسی به قانون گذاری های دیگر در مورد فریب در هوش مصنوعی در سطح ایالتی به لینک ذیل مراجعه شود:
https://www.ncsl.org/technology-and-communication/deceptive-audio-or-visual-media-deepfakes-2024-legislation

5 U.S. Congress. (2020). H.R.6216 - National Artificial Intelligence Initiative Act of 2020, 116th Congress (2019-2020). Retrieved from https://www.congress.gov/bill/116th-congress/house-bill/6216
U.S. Congress. (2023). H.R.5628 - Algorithmic Accountability Act of 2023, 118th Congress (2023-2024). Retrieved from https://www.congress.gov/bill/118th-congress/house-bill/5628/text
U.S. Congress. (2023). S.3050 - Artificial Intelligence Advancement Act of 2023, 118th Congress (2023-2024). Retrieved from https://www.congress.gov/bill/118th-congress/senate-bill/3050
U.S. Senate. (2023). S.1865 - Transparent Automated Governance Act. https://www.congress.gov/bill/118th-congress/senate-bill/1865/text

مسائل محلی و سرعت عمل آنها بوده است که نمونه‌های موفق آن در کالیفرنیا و تگزاس قابل مشاهده است. در مقابل، سطح فدرال به دلیل پیچیدگی‌های قانونگذاری ملی، نیاز به اجماع و کندی فرآیندها، عملکرد محدودتری در این زمینه داشته است.

برخلاف ایالت متحده، اتحادیه اروپا در سال‌های اخیر با درک اهمیت و حساسیت موضوع، رویکردی پیشگامانه در تدوین، تصویب و پیاده‌سازی چارچوب‌های حقوقی جامع و منسجم در حوزه فناوری هوش‌مصنوعی اتخاذ کرده است. این چارچوب‌ها به‌طور ویژه بر مقابله با چالش‌های هوش‌مصنوعی ازجمله پدیده فریب در هوش‌مصنوعی، با تمرکز خاص بر دیپ فیک‌ها، تاکید دارد. بدین ترتیب اتحادیه اروپا با ایجاد چارچوب حقوقی منسجم، شامل سه قانون کلیدی (قانون هوش‌مصنوعی[1]، قانون خدمات دیجیتالی[2] و مقررات اروپایی حفاظت از داده[3]) به مقابله با فریب هوشمند می‌پردازد. این اسناد حقوقی دارای الزامات دقیقی هستند و با ایجاد تعهدات حقوقی الزام‌آور، چارچوب نظارتی قدرتمندی را برای حفاظت از حقوق شهروندان و مقابله با سوءاستفاده‌های احتمالی فراهم می‌کنند.

از مهمترین این موارد، قانون هوش‌مصنوعی اتحادیه اروپا مصوب ۲۰۲٤ است که الزامات خاصی در این مورد دارد[4]. این قانون سیستم‌های هوش‌مصنوعی را بر اساس سطح خطر طبقه‌بندی می‌کند و محدودیت‌های خاصی برای تولید و انتشار دیپ فیک در نظر گرفته است. به موجب این قانون، سیستم‌های هوش‌مصنوعی که زمینه [۱] خطر غیرقابل‌قبول [۲] خطر بالا [۳] خطر محدود و [٤] خطر کم یا حداقلی را ایجاد می‌کنند[5]، با الزمات حقوقی متفاوتی مواجهه هستند. در سطح اول، سیستم‌های با خطر غیرقابل‌قبول قرار دارند که دارای احتمال و شدت آسیب بسیار بالا هستند. استفاده از این سیستم‌ها بر اساس ماده ۵ قانون، کاملاً ممنوع هستند و شامل مواردی مانند امتیازدهی اجتماعی و تشخیص چهره بدون هدف می‌شوند. در سطح دوم، سیستم‌های پرخطر قرار می‌گیرند که احتمال آسیب بالایی دارند اما با رعایت الزامات سخت‌گیرانه مجاز به فعالیت هستند که شامل هوش‌مصنوعی در تجهیزات پزشکی، زیرساخت‌های حیاتی، اجرای قانون و آموزش می‌شوند. سطح سوم به سیستم‌های با خطر محدود اختصاص دارد که خطر متوسطی داشته و با رعایت شفافیت می‌توانند فعالیت کنند و شامل چت‌بات‌ها، هوش‌مصنوعی مولد و دسته‌بندی زیست سنجی در موارد غیرحساس می‌شوند. در نهایت، سطح چهارم مربوط به سیستم‌های با خطر حداقلی یا بدون خطر است که نیازی به رعایت الزامات خاصی ندارند. این نوع، ضمنی[6] است و به صراحت توسط هیچ یک از مواد قانون اشاره نشده و شامل هوش‌مصنوعی در سرگرمی و سیستم‌های عمومی بدون کاربردهای حیاتی می‌شوند و تنها رعایت داوطلبانه کدهای عملیاتی برای آنها توصیه می‌شود. این دسته‌بندی چهارگانه به خوبی نشان می‌دهد که قانون‌گذار اروپایی با رویکردی متناسب با سطح خطر، الزامات حقوقی مختلفی را برای کنترل و نظارت بر سیستم‌های هوش‌مصنوعی در نظر گرفته است [۲۷].

به علاوه به موجب این قانون «دیپ فیک عبارت است از محتوای تصویری، صوتی یا ویدیویی تولید شده یا دستکاری شده توسط هوش‌مصنوعی که به شکل واقع‌گرایانه‌ای شبیه‌سازی شده و قابلیت تقلید افراد، اشیاء، مکان‌ها، ماهیت‌ها یا رویدادهای موجود را دارد، به نحوی که برای مخاطب به‌صورت اصیل و حقیقی جلوه می‌کند و قابلیت تشخیص جعلی بودن آن برای انسان دشوار است». همچنین ، تولیدکنندگان محتوای هوش‌مصنوعی که به تولید یا دستکاری

1 Artificial Intelligence Act AI Act

2 Digital Services Act (DSA)

3 General Data Protection Regulation (GDPR)

4 Artificial Intelligence Act
European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union. Retrieved from https://eur-lex.europa.eu/eli/reg/2024/1689/oj

5 (i) unacceptable risk, (ii) high risk, (iii) limited risk, and (iv) low or minimal risk

6 Implied

تصویر، صدا یا ویدیو می‌پردازند، باید مصنوعی بودن محتوای تولید شده را افشا کنند. این الزام، شامل تمامی محتواهایی می‌شود که به‌طور قابل توجهی شبیه به افراد، اشیاء، مکان‌ها یا رویدادهای واقعی هستند و ممکن است برای مخاطب؛ واقعی به نظر برسند. البته در مواردی که این محتوا برای اهداف قانونی مانند تشخیص، پیشگیری یا تحقیق جرائم استفاده می‌شود، این الزام اعمال نمی‌شود [۲۷].

بر اساس این قانون، تعهداتی نسبت به شفافیت برای ارائه‌دهندگان و توسعه دهندگان سیستم‌های هوش‌مصنوعی تعیین شده است. ارائه‌دهندگان این سیستم‌ها موظف هستند به‌صورت شفاف به کاربران درباره تعامل با سیستم هوش‌مصنوعی اطلاع‌رسانی کنند. همچنین باید تمامی خروجی‌های تولید شده توسط هوش‌مصنوعی، اعم از صوت، تصویر، ویدیو و متن را با برچسب‌های مناسب و قابل خواندن توسط ماشین نشانه‌گذاری (برچسب گذاری[1]) نمایند. علاوه بر این، ارائه راهکارهای فنی موثر، قابل همکاری و قابل اعتماد از دیگر وظایف آنها است. از سوی دیگر، توسعه دهندگان سیستم‌های هوش‌مصنوعی نیز مکلف هستند در صورت استفاده از سیستم‌های تشخیص احساسات و طبقه‌بندی زیست‌سنجی، این موضوع را به اطلاع کاربران برسانند. همچنین آنها موظف به افشای شفاف هرگونه محتوای دیپ فیک، شامل تصاویر، صداها و ویدیوهای دستکاری شده هستند. این مجموعه الزامات با هدف ایجاد شفافیت بیشتر و افزایش اعتمادپذیری در حوزه کاربرد فناوری هوش‌مصنوعی تدوین و اجرایی شده است [۲۷]. با توجه به اشاره به «برچسب گذاری» به‌عنوان الزامی مهم برای شفافیت و آگاهی‌بخشی به کاربران؛ توضیح مختصری در این خصوص مفید است. برچسب‌ها مورد اشاره باید دارای ویژگی‌های خاصی باشند تا اثربخشی لازم را داشته باشند. آنها باید به وضوح قابل مشاهده باشند و در محل مناسبی مانند گوشه تصویر یا ابتدای ویدیو قرار گیرند، همچنین این برچسب‌ها باید غیرقابل حذف بوده و به زبانی قابل فهم برای کاربر نوشته شوند. در برچسب‌ها باید اطلاعات مهمی مانند تولید محتوا توسط هوش‌مصنوعی، نوع فناوری استفاده شده، نام تولیدکننده یا منتشرکننده و تاریخ تولید محتوا درج شود. پلتفرم‌ها باید از الگوریتم‌های تشخیص محتوی مصنوعی استفاده کنند و سیستم خودکار برچسب‌گذاری را پیاده‌سازی نمایند. همچنین باید ابزارهای گزارش‌دهی در اختیار کاربران قرار دهند و سوابق محتوای مصنوعی را ذخیره کنند. کاربران حق دارند محتوای بدون برچسب را گزارش کنند، به اطلاعات بیشتر درباره نحوه تولید محتوا دسترسی داشته باشند و به برچسب‌گذاری نادرست اعتراض کنند. پلتفرم‌های بزرگی مانند متا، یوتوب و توتیر [ایکس فعلی] هر کدام سیستم‌های خاص خود را برای برچسب‌گذاری محتوی مصنوعی، پیاده‌سازی کرده‌اند. این سیستم مزایای متعددی مانند افزایش شفافیت برای کاربران، کاهش گمراه‌سازی، حفاظت از حقوق افراد و امکان تصمیم‌گیری آگاهانه را به همراه دارد.

چارچوب حقوقی دیگر در اتحادیه اروپا، قانون خدمات دیجیتالی[2] است که در سال ۲۰۲۲ تصویب و از ۲۰۲٤ اجرایی شده است. این بستر، مجموعه‌ای از الزامات حقوقی است که به شفافیت، مسئولیت‌پذیری و مدیریت محتوای آنلاین می‌پردازد. اگرچه این قانون مستقیماً به پدیده دیپ‌فیک اشاره نکرده، اما چارچوب جامعی برای مقابله با محتوای گمراه‌کننده و دستکاری‌شده ارائه می‌دهد. این مقررات با تمرکز بر پلتفرم‌های بزرگ آنلاین و موتورهای جستجو، مجموعه‌ای از تعهدات و مسئولیت‌ها را تعریف می‌کند. به‌عنوان نمونه ، متصدیان پلتفرم‌ها را از نظارت کلی بر محتوا معاف دانسته است، اما باید یک نقطه ارتباط رسمی با مقامات داشته باشند. همچنین آنها را ملزم می‌کند سیاست‌های نظارتی خود را شفاف کرده و گزارش‌های منظم ارائه دهند. نکته مهم

[1] labelling

[2] The Digital Services Act (DSA)
European Parliament & Council of the European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance)
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065

در ماده ۳۵ است که پلتفرم‌های بزرگ را موظف به کاهش خطرات سیستمی مانند انتشار اطلاعات نادرست می‌کند. در شرایط بحرانی نیز ، کمیسیون اروپا می‌تواند مداخله کند. برای تشویق خودتنظیمی نیز، تدوین منشورهای رفتاری را پیشنهاد می‌دهد و در نهایت امکان جریمه متخلفان را فراهم می‌کند. این چارچوب حقوقی با ترکیب رویکردهای پیشینی و پسینی و تأکید بر شفافیت و مسئولیت‌پذیری، گامی مهم در تنظیم‌گری فضای دیجیتال محسوب می‌شود[۳٤] .

چارچوب حقوقی دیگر؛ مقررات اروپایی حفاظت از داده[1] مصوب ۲۰۱٦ است. اگرچه این قانون مستقیماً به پدیده دیپ‌فیک اشاره نکرده لیکن ارتباط این مقررات با فریب هوشمند را می‌توان در دوجنبه بیان نمود. نخست اینکه به‌طور کلی مقررات اروپایی حفاظت از داده، چارچوب اصلی برای محافظت از حقوق دیجیتالی افراد را تشکیل می‌دهد. این بستر حقوقی، با تأکید بر وجود مبانی حقوقی برای پردازش، استفاده از داده‌های شخصی برای تولید دیپ فیک را محدود می‌کنند. همچنین در حوزه مسئولیت‌ها نیز، اشخاص پردازش کننده داده، موظف به پیاده‌سازی اقدامات امنیتی و فنی مناسب هستند. این امر شامل ارزیابی خطرات قبل از استفاده از فناوری‌های دیپ فیک و مسئولیت‌پذیری در قبال نشت احتمالی داده‌های شخصی می‌شود. از جنبه ایی دیگر و به‌طور خاص نیز بر اساس این قانون، اصول حاکم بر پردازش داده‌ها در تولید دیپ‌فیک نیز قابل استفاده است که بر اساس آن، هرگونه استفاده از داده‌های شخصی برای ساخت دیپ‌فیک باید مطابق با اصول کلی حاکم بر پردازش باشد. همچنین به حقوق اشخاص موضوع داده اشاره دارد که به افراد این حق را می‌دهد تا از استفاده تصاویر و صدای خود در دیپ‌فیک مطلع شده و امکان کنترل بر آن را داشته باشند، در ادامه تعهدات پردازش‌کنندگان داده را مشخص می‌کند که سازندگان دیپ‌فیک را ملزم به رعایت اصول حفاظتی و امنیتی می‌نماید. در نهایت نیز ضمانت اجرای تخلفات را تعیین کرده و جریمه‌های سنگینی برای نقض مقررات در نظر گرفته است که در خصوص فریب هوش‌مصنوعی نیز قابل استفاده‌اند.

ج) رویکرد قانونی نظام حقوقی ایران

ایران در حال حاضر، قانون خاصی در زمینه فریب هوش‌مصنوعی ندارد و با خلاء قانونی مواجهه است. همچنین، قانون کلی نیز در حوزه هوش‌مصنوعی در کشور تصویب نشده است. از سوی دیگر، قانون جامع حمایت از داده‌های شخصی نیز هنوز به تصویب نرسیده است. بدین ترتیب در چنین شرایطی، برای رسیدگی به موارد مرتبط با فریب هوش‌مصنوعی، باید به قوانین مرتبط موجود مانند قانون جرایم رایانه‌ای[۲] مصوب ۱۳۸۸ و برخی از مصوبات شورای عالی فضای مجازی در مورد حریم خصوصی و داده‌های کاربران[۳]؛ مراجعه کرد. این چارچوب‌های مرتبط می‌توانند به‌صورت غیرمستقیم برای رسیدگی به موارد فریب هوش‌مصنوعی مورد استفاده قرار گیرند، هرچند که کافی و جامع نیستند. البته ناگفته نماند در ایران نسبت به تنظیم گری هوش‌مصنوعی تاکنون دو اقدام سندی و یک اقدام نهادی انجام شده‌است. این اسناد، «سند ملی هوش‌مصنوعی» مصوبه شورای عالی انقلاب فرهنگی در تیرماه ۱٤۰۳ و «طرح ملی هوش‌مصنوعی» است که در تاریخ ۲۹ مهر ۱٤۰۳ در مجلس شورای اسلامی اعلام وصول شده و هدف آن ایجاد چارچوب قانونی برای توسعه و تنظیم‌گری هوش‌مصنوعی در کشور است و قرار است زیرساخت‌های حقوقی لازم برای تحقق هدف قرارگیری ایران در میان ۱۰ کشور برتر حوزه هوش‌مصنوعی را فراهم کند. همچنین «سازمان ملی هوش‌مصنوعی مصوب در جلسه شورای عالی انقلاب فرهنگی در خرداد ۱٤۰۳ نیز اقدام نهادی ایران در خصوص هوش‌مصنوعی است. این موارد بیشتر جنبه سیاست‌گذاری کلان – تا برنامه اجرایی– دارند و فاقد جزئیات اجرایی مشخص و الزامات جقوقی خاص هستند. بدین ترتیب هم به

[1] General Data Protection Regulation (GDPR)
https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[۲] در این خصوص، مواد مقرر در فصول دوم و سوم در مورد کلاهبرداری رایانه ای و جعل رایانه ای قابل استفاده هستند.

[۳] بررسی کامل مصوبات از لینک ذیل قابل دسترسی است:
https://dotic.ir/cat/120/1

دلیل وضعیت اعتبار و هم ضعف در الزامات حقوقی؛ قابل استفاده نیستند. از سویی دیگر نظام حقوقی ایران در حوزه حفاظت از داده‌های شخصی نیز با خلأهای قانونی عمیقی مواجه است. البته ایران با ارائه پیش نویس های قانونی –طرح حمایت و حفاظت از داده و اطلاعات شخصی مرکز پژوهش‌های مجلس شورای اسلامی (۱۴۰۰) و طرح حفاظت از داده شخصی مرکز پژوهش‌های مجلس شورای اسلامی (۱۴۰۳)– در این خصوص گامی برداشته‌است.

با توجه به وضعیت کنونی نظام حقوقی ایران، ضرورت دارد راهکاری مناسبی برای بهبود شرایط موجود شناسایی شود. راهکار پیشنهادی که در ادامه ارائه خواهد شد، می‌تواند به‌عنوان راه‌حل‌های عملی مورد توجه قرار گیرد.

۵-۲ تنظیم‌گری دقیق و هدفمند متناسب با فناوری

در حوزه تنظیم‌گری هوش‌مصنوعی، ضرورت ایجاد یک چارچوب جامع که ابعاد مختلف حقوقی، فنی، الگوبرداری موفق و همکاری‌های بین‌المللی را پوشش دهد، بیش از پیش احساس می‌شود. در بعد حقوقی، تصویب قوانین منعطف و پویا که توانایی همگام‌سازی با تحولات سریع فناوری را داشته باشند، همراه با تعیین استانداردهای الزام‌آور و مکانیسم‌های نظارتی قوی، اساس این چارچوب را تشکیل می‌دهد [۳۶،۳۵]. از سوی دیگر در بعد فنی، پیاده‌سازی سیستم‌های واترمارکینگ[1] پیشرفته و توسعه الگوریتم‌های هوشمند تشخیص محتوای جعلی، همراه با استقرار زیرساخت‌های امنیتی قدرتمند و مکانیسم‌های احراز هویت چندلایه، ضامن اجرای موفق قانون خواهد بود [۳۷]. این سیستم‌ها باید به‌گونه‌ای طراحی شوند که ضمن حفظ نوآوری، از سوءاستفاده‌های احتمالی جلوگیری کنند [۳۸]. به‌علاوه بهره‌مندی از تجربه موفق نظام‌هایی مانند اتحادیه اروپا در زمینه تنظیم‌گری فناوری‌های نوین می‌تواند به‌عنوان الگویی کارآمد – البته با در نظر گرفتن شرایط و اقتضائات بومی هر کشور و صرفاً به تقلید محض محدود نشود. – مورد استفاده قرار گیرد [۳۹]. در بعد همکاری‌های بین‌المللی، ایجاد شبکه‌های همکاری میان نهادهای تنظیم‌گر و تبادل مستمر اطلاعات و تجربیات، همراه با مشارکت فعال در تدوین کنوانسیون‌های بین‌المللی و توسعه پروتکل‌های مشترک مقابله با تهدیدات، نقشی حیاتی در موفقیت این چارچوب تنظیمی ایفا می‌کند .

۶- نتیجه‌گیری

توسعه هوش‌مصنوعی در سال‌های اخیر، نظام‌های حقوقی جهان را با چالش‌های بنیادین مواجه ساخته است. ظهور پدیده‌ی فریبِ هوشمند، ضرورت توجه به چگونگی مقابله با این امر و بالتبع تنظیم‌گری دقیق حوزه هوش‌مصنوعی را بیش‌ازپیش نمایان کرده است. در این راستا افزایش چشمگیر انتشار اطلاعات نادرست توسط سیستم‌های هوش‌مصنوعی و طرح دعاوی متعدد افترا علیه شرکت‌های فناوری، نشان‌دهنده ضعف و ناکارآمدی چارچوب‌های حقوقی کنونی در مواجهه با این چالش‌های نوظهور است. بدین جهت در پاسخ به این چالش‌ها، نظام‌های حقوقی پیشرو به دنبال راهکارهایی موثر برای تنظیم‌گری هوش‌مصنوعی برآمدند. مسأله اصلی در این میان، یافتن چارچوب حقوقی مناسبی است که بتواند ضمن حفظ حقوق اساسی افراد، مسئولیت‌های ناشی از فریب و سوءاستفاده از سیستم‌های هوش‌مصنوعی را به‌درستی تعیین و توزیع نماید. در این راستا، تحولات قانونگذاری اخیر، به‌ویژه در حوزه اتحادیه اروپا، نشان‌دهنده رویکرد جدیدی در مواجهه با این چالش‌هاست. در این مسیر به جای پذیرش نظریاتی همچون «شخصیت حقوقی مشتق، به سمت اتخاذ رویکرد «مسئولیت زنجیره‌ای» حرکت شده است. این رویکرد که مبتنی بر تقسیم نظام‌مند مسئولیت‌ها میان ذینفعان مختلف است، چارچوب منسجم و کاربردی را برای مواجهه با چالش‌های حقوقی هوش‌مصنوعی فراهم می‌آورد. این امر بدین دلیل است که ساختار مسئولیت نسبی در این رویکرد، با تعیین دقیق نقش و مسئولیت هر یک از اجزای زنجیره ارزش؛ همچنین امکان انتقال مسئولیت‌ها، پاسخگوی نیازهای نوظهور در این حوزه است. به علاوه، رویکرد مبتنی بر خطر در تعیین

[1] Watermarking
نوعی تکنیک دیجیتال است که برای محافظت از محتوا و تأیید اصالت آن استفاده می‌شود.

سطوح مسئولیت، انعطاف‌پذیری لازم برای مواجهه با پیچیدگی‌های فزاینده سیستم‌های هوش‌مصنوعی را تأمین می‌کند. بنابراین به نظر می‌رسد مسیر آینده تنظیم‌گری هوش‌مصنوعی، نه در اعطای شخصیت حقوقی به این سیستم‌ها، بلکه در توسعه و تکامل نظام مسئولیت و تقویت سازوکارهای نظارتی متناسب با سطوح مختلف خطر نهفته است.

این رویکرد نوین در تنظیم‌گری هوش‌مصنوعی که بر پایه مسئولیت زنجیره‌ای استوار است، می‌تواند الگوی مناسبی برای سایر نظام‌های حقوقی در حال توسعه باشد. به‌ویژه برای کشورهایی که هنوز چارچوب قانونی مشخصی برای تنظیم هوش‌مصنوعی تدوین نکرده‌اند. البته کشورهایی که با فقدان چارچوب حقوقی منسجم برای هوش‌مصنوعی مواجهه هستند – مانند نظام حقوقی ایران– باید به‌طور خاص در مسیر استقرار نظام جامع حقوقی– نظارتی نسبت به هوش‌مصنوعی، گام بردارند. این مهم مستلزم اتخاذ رویکردی نظام‌مند و چندمرحله‌ای است. در مرحله نخست، تأسیس کارگروهی متشکل از متخصصان حوزه‌های حقوق، فناوری و سیاست‌گذاری ضروری است. این کارگروه موظف است ضمن مطالعه تطبیقی نظام‌های حقوقی پیشرو –علی‌الخصوص قانون هوش‌مصنوعی اتحادیه اروپا– به احصاء خلاهای قانونی موجود در نظام حقوقی کشور بپردازد. بدین ترتیب در مرحله تقنینی، تدوین پیش‌نویس قانونی مستلزم رعایت اصولی است که در درجه نخست، طبقه‌بندی مخاطرات را شامل شود؛ به‌نحوی‌که تفکیک سطوح خطر در حوزه هوش‌مصنوعی، شناسایی و دسته‌بندی انواع فریب مصنوعی و تعیین آستانه‌های خطر برای هر طبقه را در برگیرد. در گام بعدی، تبیین حدود مسئولیت‌های قانونی ضرورت می‌یابد که این امر از طریق تعیین مسئولیت‌های مدنی و کیفری اطراف درگیر، تدوین ضوابط احراز رابطه سببیت و پیش‌بینی نظام جبران خسارت محقق می‌گردد. همچنین در راستای تضمین اجرای قانون، تعیین ضمانت‌های اجرایی که مشتمل بر وضع مجازات متناسب با شدت تخلف، پیش‌بینی اقدامات تأمینی و پیشگیرانه و تعیین سازوکارهای نظارت و پایش مستمر باشد؛ از اهمیت ویژه‌ای برخوردار است. به موازات فرآیند قانون‌گذاری، استقرار سازو کارهای نظارتی موقت با ویژگی‌های خاص ضروری است که باید از انعطاف‌پذیری کافی در مواجهه با تحولات فناورانه برخوردار بوده، قابلیت اجرای فوری و کم‌هزینه داشته و امکان اصلاح و تکمیل بر اساس بازخوردهای عملیاتی را فراهم آورد. از سویی دیگر در افق بلندمدت، توسعه زیرساخت‌های فنی و حقوقی از طریق دو محور اصلی باید پیگیری شود. این موارد تربیت نیروی انسانی متخصص و دوم، تقویت دیپلماسی حقوقی–فناورانه است که از طریق عضویت فعال در نهادهای بین‌المللی مرتبط، مشارکت در تدوین استانداردهای جهانی و تبادل تجربیات با کشورهای پیشرو محقق می‌گردد.

تعارض منافع

نویسندگان تعهد می‌کنند که هیچ تعارض منافعی در این مقاله وجود نداشته‌است.

References

[1] Fischer, K. A. (2023). **Reflective Linguistic Programming (RLP): A Steppingstone In Socially-Aware AGI (Socialagi)** (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2305.12647

[2] Sodian, B., Taylor, C., Harris, P. L., & Perner, J. (1991). **Early Deception And The Child's Theory Of Mind: False Trails And Genuine Markers**. *Child Development*, 62(3), 468. https://doi.org/10.2307/1131124

[3] Chandler, M., Fritz, A. S., & Hala, S. (1989). **Small-Scale Deceit: Deception As A Marker Of Two-, Three-, And Four-Year-Olds' Early Theories Of Mind**. *Child Development*, 60(6), 1263. https://doi.org/10.2307/1130919

[4] Fatemi, M. Y., Suttle, W. A., & Sadler, B. M. (2024). **Deceptive Path Planning Via Reinforcement Learning With Graph Neural Networks** (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2402.06552

[5] Liu, Z., Yang, Y., Miller, T., & Masters, P. (2021). **Deceptive Reinforcement Learning For Privacy-Preserving Planning**. *arXiv*. https://doi.org/10.48550/ARXIV.2102.03022

[6] Guo, L. (2024). **Unmasking The Shadows Of AI: Investigating Deceptive Capabilities In Large Language Models** (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2403.09676

[7] Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., Roller, S., … Zijlstra, M. (2022). **Human-Level Play In The Game Of Diplomacy By Combining Language Models With Strategic Reasoning**. *Science*, 378(6624), 1067–1074. https://doi.org/10.1126/science.ade9097

[8] Hendrycks, D., Mazeika, M., & Woodside, T. (2023). **an Overview of Catastrophic AI Risks** (Version 6). *arXiv*. https://doi.org/10.48550/ARXIV.2306.12001

[9] Dogra, A., Pillutla, K., Deshpande, A., Sai, A. B., Nay, J., Rajpurohit, T., Kalyan, A., & Ravindran, B. (2024). **Deception In Reinforced Autonomous Agents** (Version 2). *arXiv*. https://doi.org/10.48550/ARXIV.2405.04325

[10] Ward, F. R., Belardinelli, F., Toni, F., & Everitt, T. (2023). **Honesty Is The Best Policy: Defining And Mitigating AI Deception** (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2312.01350

[11] Zhong, H., O'Neill, E., & Hoffmann, J. A. (2024). **Regulating AI: Applying Insights From Behavioural Economics And Psychology To The Application Of Article 5 Of The EU AI Act**. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18), 20001–20009. https://doi.org/10.1609/aaai.v38i18.29977

[12] Alanazi, S., Asif, S., & Moulitsas, I. (2024a). **Examining The Societal Impact And Legislative Requirements Of Deepfake Technology: A Comprehensive Study**. *International Journal of Social Science and Humanity*. https://doi.org/10.18178/ijssh.2024.14.2.1194

[13] Faraoni, S. (2023). **Persuasive Technology And Computational Manipulation: Hypernudging Out Of Mental Self-Determination**. *Frontiers in Artificial Intelligence*, 6, 1216340. https://doi.org/10.3389/frai.2023.1216340

[14] Farid, H. (2022). **Creating, Using, Misusing, And Detecting Deep Fakes**. *Journal of Online Trust and Safety*, 1(4). https://doi.org/10.54501/jots.v1i4.56

[15] Castelfranchi, C. & Yao-Hua Tan. (2001). **The Role Of Trust And Deception In Virtual Societies.** *Proceedings Of The 34th Annual Hawaii International Conference on System Sciences*, 8. https://doi.org/10.1109/HICSS.2001.927042

[16] Li, L., Ma, H., Kulkarni, A. N., & Fu, J. (2023). **Dynamic Hypergames For Synthesis Of Deceptive Strategies With Temporal Logic Objectives**. *IEEE Transactions on Automation Science and Engineering*, 20(1), 334–345. https://doi.org/10.1109/TASE.2022.3150167

[17] Savas, Y., Verginis, C. K., & Topcu, U. (2022). **Deceptive Decision-Making Under Uncertainty.** *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5), 5332–5340. https://doi.org/10.1609/aaai.v36i5.20470

[18] Sarkadi, Ş., Panisson, A. R., Bordini, R. H., McBurney, P., Parsons, S., & Chapman, M. (2019). **Modelling Deception Using Theory Of Mind In Multi-Agent Systems**. *AI Communications*, 32(4), 287–302. https://doi.org/10.3233/AIC-190615

[19] Hamann, H., Khaluf, Y., Botev, J., Divband Soorati, M., Ferrante, E., Kosak, O., Montanier, J.-M., Mostaghim, S., Redpath, R., Timmis, J., Veenstra, F., Wahby, M., & Zamuda, A. (2016). **Hybrid Societies: Challenges And Perspectives In The Design Of Collective Behavior In Self-Organizing Systems**. *Frontiers in Robotics and AI*, 3. https://doi.org/10.3389/frobt.2016.00014

[20] Huang, L., & Zhu, Q. (2022). **A Dynamic Game Framework For Rational And Persistent Robot Deception With An Application To Deceptive Pursuit-Evasion**. *IEEE Transactions on Automation Science and Engineering*, 19(4), 2918–2932. https://doi.org/10.1109/TASE.2021.3097286

[21] Zhan, X., Xu, Y., Abdi, N., Collenette, J., Abu-Salma, R., & Sarkadi, S. (2024**). Banal Deception Human-AI Ecosystems: A Study Of People's Perceptions Of LLM-Generated Deceptive Behaviour** (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2406.08386

[22] Anderson, D., Stephenson, M., Togelius, J., Salge, C., Levine, J., & Renz, J. (2018). **Deceptive Games**. In K. Sim & P. Kaufmann (Eds.), **Applications Of Evolutionary Computation** (Vol. 10784, pp. 376–391). *Springer International Publishing*. https://doi.org/10.1007/978-3-319-77538-8_26

[23] Ornik, M., & Topcu, U. (2018). **Deception In Optimal Control**. *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 821–828. https://doi.org/10.1109/ALLERTON.2018.8635871

[24] Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., Righetti, L., & Saunders, W. (2021). **Truthful AI: Developing And Governing AI That Does Not Lie** (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2110.06674

[25] Tarsney, C. (2025). **Deception And Manipulation In Generative AI**. *Philosophical Studies*. https://doi.org/10.1007/s11098-024-02259-8

[26] Hasibuan, R. H., Aurelya Jessica Rawung, Denisha M. D. Paranduk, & Fidel Jeremy Wowiling. (2024). **Artificial Intelligence In The Auspices Of Law: A Diverge Perspective**. *Mimbar Hukum*, 36(1), 111–140. https://doi.org/10.22146/mh.v36i1.10827

[27] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 **(Artificial Intelligence Act)** Text with EEA relevance. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[28] Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W.-G. (2021). **Framing Governance For A Contested Emerging Technology: Insights From AI Policy**. *Policy and Society*, 40(2), 158–177. https://doi.org/10.1080/14494035.2020.1855800

[29] Cooper, A. F., Lee, K., Grimmelmann, J., Ippolito, D., Callison-Burch, C., Choquette-Choo, C. A., Mireshghallah, N., Brundage, M., Mimno, D., Choksi, M. Z., Balkin, J. M., Carlini, N., De Sa, C., Frankle, J., Ganguli, D., Gipson, B., Guadamuz, A., Harris, S. L., Jacobs, A. Z., … Zeide, E. (2023). **Report Of The 1st Workshop On Generative AI And Law** (Version 3). *arXiv*. https://doi.org/10.48550/ARXIV.2311.06477

[30] Afgiansyah, A. (2023). **Artificial Intelligence Neutrality: Framing Analysis Of GPT Powered-Bing Chat And Google Bard**. *Jurnal Riset Komunikasi*, 6(2), 179–193. https://doi.org/10.38194/jurkom.v6i2.908

[31] Pataranutaporn, P., Archiwaranguprok, C., Chan, S. W. T., Loftus, E., & Maes, P. (2024). **Synthetic Human Memories: AI-Edited Images And Videos Can Implant False Memories And Distort Recollection** (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2409.08895

[32] **Deceptive Audio Or Visual Media ("Deepfakes") 2024 Legislation**. (n.d.). Retrieved February 12, 2025, from https://www.ncsl.org/technology-and-communication/deceptive-audio-or-visual-media-deepfakes-2024-legislation

[33] Alanazi, S., Asif, S., & Moulitsas, I. (2024b). **Examining The Societal Impact And Legislative Requirements Of Deepfake Technology: A Comprehensive Study**. *International Journal of Social Science and Humanity*. https://doi.org/10.18178/ijssh.2024.14.2.1194

[34] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC **(Digital Services Act)** (Text with EEA relevance) https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

[35]Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: **A Survey Of Examples, Risks, And Potential Solutions**. *Patterns*, 5(5), 100988. https://doi.org/10.1016/j.patter.2024.100988

[36] Kim, T. W., Tong, Lu, Lee, K., Cheng, Z., Tang, Y., & Hooker, J. (2021). **When Is It Permissible For Artificial Intelligence To Lie? A Trust-Based Approach** (Version 2). *arXiv*. https://doi.org/10.48550/ARXIV.2103.05434

[37] Fartash, K., Kheiri, E. and Baramaki, T. (2024). **Providing A Framework For AI Infrastructure In Iran, With A Focus On Service Providers And Service Aggregators Of AI**. *Journal of Science and Technology Policy*, 17(3), 9-25. doi: 10.22034/jstp.2025.11771.1815 **(In Persian)**

[38] Kung, H.-W., Varanka, T., Saha, S., Sim, T., & Sebe, N. (2024). **Face Anonymization Made Simple** (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2411.00762

[39] Carvalko, J. R. (2024). **Generative AI, Ingenuity, And Law**. *IEEE Transactions on Technology and Society*, 5(2), 169–182. https://doi.org/10.1109/TTS.2024.3413591

[40] Shahbazinia, M. and Zolghadr, M. J. (2024). **Recognizing Artificial Intelligence (AI) As A Legal Person: Providing A Policy Proposal To The Iranian Legislator**. *Journal of Science and Technology Policy*, 17(3), 41-52. doi: 10.22034/jstp.2025.11778.1819 **(In Persian)**