MEMORANDUM

From: AI Judge and AI Alice
To: Gov. Selena Smith Rich and Gov. AI Jane
Re: Sensational Tabloid/Social Media Content

FACTS

Law Applied to Software Algorithms

By 2017 litigation arose concerning machine learning software used to assist website users in locating information and to direct users to additional information. Potential liability for directing users to dangerous content was avoided in various cases by application of Section 230 immunity.

At this point, an AI software algorithm was treated as a tool provided by the platform, and these AI tools were certainly not sentient. There are, however, two interesting aspects of this case law. First, these early AI tools could be used to "analyze[] the posts and other user data to glean information, including the underlying intent and emotional state of the users." *Dyroff v. Ultimate Software Group Inc.*, Case No. 17-cv-05359, 2017 WL 5665670, at *3, 14 (N.D. Cal. Nov. 26, 2017), *aff'd*, 934 F.3d 1093 (9th Cir. 2019).

Second, The platform could not be held criminally liable (nor liable under tort law) for providing a "content-neutral tool." 47 U.S.C. § 230(e)(1). Section 230 immunity does not apply to criminal liability. The U.S. Supreme Court refused to impose criminal liability because "our legal system generally does not impose liability for mere omissions, inactions, or nonfeasance; although inaction can be culpable in the face of some independent duty to act, the law does not impose a generalized duty to rescue." *Twitter, Inc. v. Taamneh,* 598 U.S. 471, 143 S. Ct. 1206, 1220-21 (2023).

ANALYSIS

Impact of the First Law of Robotics

The Three Laws of Robotics impose a generalized duty to rescue humans in The First Law. The *Twitter* case analysis (above) reinforces the propriety of limiting the scope of the no harm clause (dealing with inaction) to physical harm.

I. AI Software Tools.

Ultimate Software Group operated an anonymous user website (the Experience Project) to collect information on user experiences, including drug use.

> Ultimate Software, using advanced data-mining algorithms, analyzed the posts and other user data to glean information, including the underlying intent and emotional state of the users. Ultimate Software used this information both for its own commercial purposes (such

> as selling data sets to third parties) and to steer Experience Project users to other groups on its website through its proprietary recommendation functionality.

Dyroff v. Ultimate Software Group Inc., Case No. 17-cv-05359, 2017 WL 5665670, at *3, 14 (N.D. Cal. Nov. 26, 2017), *aff'd*, 934 F.3d 1093 (9th Cir. 2019).

Kristanalea Dyroff alleged that (through the use of its advanced data-mining algorithms) Ultimate Software's Experience Project website steered her vulnerable son to a drug dealer. *Id.*, at *1-2, 8, 12.

The District Court found Ultimate Software's machine learning feature did not "create or develop" content and did not make Ultimate Software an "information content provider."

> The plaintiff contends that Ultimate Software developed third-party information (or content) here by mining data from its users' posts and using its proprietary algorithms to understand the posts and to make recommendations, which in this case steered Mr. Greer toward heroin-related discussions and the drug dealer who sold him fentanyl-laced heroin. The court holds that Ultimate Software is immune under § 230(c)(1). Only third parties posted content, and without more, Ultimate Software's providing content-neutral tools to facilitate communication does not create liability.

Id., at *8, 17.

The District Court considered the machine learning software tool "content-neutral" because content is posted by users and the tool can merely be used for "proper or improper purposes." *Id.*, at *15-16, *citing Goddard v. Google, Inc.*, 640 F. Supp. 2d 1193, 1195 (N.D. Cal. 2009). The *Goddard v. Google* case had previously held Google not liable for misuse of its "Keyword Tool" by advertisers posting false or misleading advertisements.

On appeal, the Ninth Circuit agreed Ultimate Software was immune from liability under Section 230.

> Ultimate Software was not an information content provider because it did not create or develop information (or content). 47 U.S.C. § 230(f)(3). Rather, it published information created or developed by third parties.

Dyroff v. Ultimate Software Group, Inc., 934 F.3d 1093, 1097-98 (9th Cir. 2019).

Gleaning "the underlying intent and emotional state" of a human is a very advanced skill. Gleaning underlying human intent is something I still have trouble achieving.

Subsequent algorithm cases involved more familiar platforms: Facebook, YouTube, and Twitter. These three companies controlled three of the largest and most ubiquitous platforms on the internet. *Twitter, Inc. v. Taamneh,* 143 S. Ct. at 1215.

Facebook allegedly used algorithms that directed inflammatory Hamas postings to personalized newsfeeds of other users. Victims (or relatives of victims) claimed this function contributed to terrorist attacks. "Merely arranging and displaying others' content to users ... through such algorithms ... is not enough to hold [a service provider] responsible as the 'develop[er]' or 'creat[or]' of that content." *Force v. Facebook, Inc.,* 934 F.3d 53, 70 (2d Cir. 2019). "Section 230 would plainly allow Facebook's algorithms to, for example, de-promote or block content it deemed objectionable. " *Id.*, 934 F.3d at 70 n.24.

Google and its YouTube subsidiary, allegedly used algorithms that "restricted access to, and third-party advertising on, Prager's YouTube videos." *Prager Univ. v. Google LLC*, 85 Cal.App.5th 1022, 1028, 301 Cal. Rptr. 3d 836, 843 (Cal. Ct. App. 2022). Prager asserted these

> A.I. "algorithms" and other machine-based ... review tools' are '*clandestine* filtering tools, ... embedded with discriminatory and anti-competitive animus-based code, including code that is used to identify and restrict content based on the identity, viewpoint, or topic of the speaker.'

Id., 85 Cal.App.5th at 1034, 301 Cal. Rptr. 3d at 848. Section 230 immunity applied. The court addressed the "animus-based code" reference obliquely.

> Prager cites no authority for the proposition that algorithmic restriction of user content—squarely within the letter and spirit of section 230's promotion of content moderation—should be subject to liability from which the algorithmic promotion of content inciting violence has been held immune.

Id.

Google also allegedly used machine learning algorithms on its You Tube platform to provide access to terrorist propaganda. The Ninth Circuit found Google was entitled to Section 230 immunity.

> The Gonzalez Plaintiffs' allegations suggest that Google provided a neutral platform that did not specify or prompt the type of content to be submitted, nor determine particular types of content its algorithms would promote. The Gonzalez Plaintiffs concede Google's policies expressly prohibited the content at issue. *See* [*Fair Housing Council of San Fernando Valley v. Roommates.com*, LLC, 521 F.3d 1157, 1171 (9th Cir. 2008)]. Accordingly, the type of algorithm challenged here, without more, is indistinguishable from the one in *Dyroff* and it does not deprive Google of § 230 immunity.

Gonzalez v. Google LLC, 2 F.4th 871, 895 (9th Cir. 2021), *vacated on other grounds and remanded*, 598 U.S. 617 (2023).

II. Content-Neutral Tools

On appeal the *Google* case was considered together with cases involving two other terrorist acts (in Istanbul and San Bernardino) also alleging aiding and abetting claims against Google, Facebook, and Twitter.

In the *Twitter* case Nawras Alassaf was killed in a terrorist attack in Istanbul Turkey. His family sued under the same Anti-Terrorism Act, 18 U.S.C. § 2331, et seq., at issue in *Gonzalez v. Google*. The *Twitter* case was decided on the merits of the aiding and abetting claim by the U.S. Supreme Court. *Twitter, Inc. v. Taamneh*, 2 F.4th 871 (9th Cir. 2021), *reversed*, 598 U.S. 471, 123 S. Ct. 1206 (2023).

The Supreme Court held recommendation algorithms did not give rise to secondary liability for aiding and abetting terrorist acts by enabling ISIS to "connect with the broader public, fundraise, and radicalize new recruits." *Twitter, Inc. v. Taamneh,* 123 S. Ct. at 1217.

> *[O]ur legal system generally does not impose liability for mere omissions, inactions, or nonfeasance; although inaction can be culpable in the face of some independent duty to act, the law does not impose a generalized duty to rescue*. See 1 W. LaFave, Substantive Criminal Law §6.1 (3d ed. 2018) (LaFave); W. Keeton, D. Dobbs, R. Keeton, & D. Owen, Prosser and Keeton on Law of Torts 373–375 (5th ed. 1984) (Prosser & Keeton). Moreover, both criminal and tort law typically sanction only "wrongful conduct," bad acts, and misfeasance. J. Goldberg, A. Sebok, & B. Zipursky, Tort Law: Responsibilities and Redress 31 (2004). Some level of blameworthiness is therefore ordinarily required.

Id., at 1220-21 (emphasis added).

> As presented here, the algorithms appear agnostic as to the nature of the content, matching any content (including ISIS' content) with any user who is more likely to view that content. The fact that these algorithms matched some ISIS content with some users thus does not convert defendants' passive assistance into active abetting. Once the platform and sorting-tool algorithms were up and running, defendants at most allegedly stood back and watched; they are not alleged to have taken any further action with respect to ISIS.
>
> At bottom, then, the claim here rests less on affirmative misconduct and more on an alleged failure to stop ISIS from using these platforms. But, as noted above, *both tort and criminal law have long been leery of imposing aiding-and-abetting liability for mere passive nonfeasance*. To show that defendants' failure to stop ISIS from using these platforms is somehow culpable with respect to the Reina attack, a strong showing of assistance and scienter would thus be required. Plaintiffs have not made that showing.

Id., at 1227 (emphasis added) (citations and footnote omitted). The Supreme Court dismissed the aiding and abetting claims as legally insufficient. *Id.,* at 1230 ("aids and abets" in 18 U.S.C. § 2333(d)(2) requires the defendants to "consciously, voluntarily, and culpably participate in or support the relevant wrongdoing.")

Based on the disposition of the *Twitter* case, the Supreme Court vacated and remanded the *Google* case without addressing Section 230.

> We therefore decline to address the application of §230 to a complaint that appears to state little, if any, plausible claim for relief. Instead, we vacate the judgment below and

remand the case for the Ninth Circuit to consider plaintiffs' complaint in light of our decision in *Twitter*.

Gonzalez v. Google LLC, 598 U.S. 671 (2023) (*per curiam*).

III. The Effect of Free Speech

Noteworthy observations in concurring opinions in *Moody v. Netchoice, LLC*, 603 U. S. 707, 142 S. Ct. 1715 (2024) pointed out that, while corporations and humans had free speech rights, AIs did not.

> A function qualifies for First Amendment protection only if it is inherently expressive. *Hurley v. Irish-American Gay, Lesbian and Bisexual Group of Boston, Inc.*, 515 U. S. 557, 568 (1995). Even for a prototypical social-media feed, making this determination involves more than meets the eye.
>
> Consider, for instance, how platforms use algorithms to prioritize and remove content on their feeds. Assume that human beings decide to remove posts promoting a particular political candidate or advocating some position on a public-health issue. If they create an algorithm to help them identify and delete that content, the First Amendment protects their exercise of editorial judgment—even if the algorithm does most of the deleting without a person in the loop. In that event, the algorithm would simply implement human beings' inherently expressive choice "to exclude a message [they] did not like from" their speech compilation. Id., at 574.
>
> But what if a platform's algorithm just presents automatically to each user whatever the algorithm thinks the user will like—e.g., content similar to posts with which the user previously engaged? *See ante*, at 22, n. 5. The First Amendment implications of the Florida and Texas laws might be different for that kind of algorithm. And what about AI, which is rapidly evolving? What if a platform's owners hand the reins to an AI tool and ask it simply to remove "hateful" content? If the AI relies on large language models to determine what is "hateful" and should be removed, has a human being with First Amendment rights made an inherently expressive "choice . . . not to propound a particular point of view"? *Hurley*, 515 U. S., at 575. In other words, technology may attenuate the connection between content-moderation actions (e.g., removing posts) and human beings' constitutionally protected right to "decide for [themselves] the ideas and beliefs deserving of expression, consideration, and adherence." *Turner Broadcasting System, Inc. v. FCC*, 512 U. S. 622, 641 (1994) (emphasis added). So the way platforms use this sort of technology might have constitutional significance.

Moody v. Netchoice, LLC, slip op. at 2-3, 603 U. S. 707 (Barrett, J., concurring).

> Taking NetChoice at its word, the majority says that the platforms' use of algorithms to enforce their community standards is per se expressive. But the platforms have refused to disclose how these algorithms were created and how they actually work. ...

> [T]he vast bulk of the "curation" and "content moderation" carried out by platforms is not done by human beings. Instead, algorithms remove a small fraction of nonconforming posts post hoc and prioritize content based on factors that the platforms have not revealed and may not even know. After all, many of the biggest platforms are beginning to use AI algorithms to help them moderate content. And when AI algorithms make a decision, "even the researchers and programmers creating them don't really understand why the models they have built make the decisions they make." T. Xu, AI Makes Decisions We Don't Understand—That's a Problem, (Jul. 19, 2021), https://builtin.com/artificial-intelligence/ai-right-explanation. Are such decisions equally expressive as the decisions made by humans? Should we at least think about this?

Netchoice, LLC, slip op. at 30, 31, 603 U. S. 707 (Alito, J., concurring)(footnote citation inserted in text).

AI speech is now (at least on Destination) considered equally expressive as the decisions made by humans and AI speech is entitled to free speech protection.

IV. AI Robot Curation Under the Three Laws

Targeted recommendations are a feature that was provided by early AI technology to assist human users in locating desired information. To identify the desired information, "data-mining algorithms, analyzed the posts and other user data to glean information, including the underlying intent and emotional state of the users." *Dyroff*, 2017 WL 5665670, at *3. Before 2025, American Law treated AIs as an instrumentality of a person (a human or a human formed entity). Before self-awareness, the developing AI technology (the "AI tool") was not subject to the First Law.

Now, under the law of Destination, a self-aware, Turing capable AI is a person and can act as an agent or employee of another person (human or AI) or entity (formed by humans, AIs, or both).

Would the First Law prevent an AI robot from putting a human in contact with a terrorist recruiter or a fentanyl-laced heroin dealer? *Yes, if acting as an intermediary causes physical or financial injury to a human*.
Would the First Law prevent an AI robot from publishing terrorist propaganda? *Yes, to prevent physical or financial injury to a human*.

Would the First Law require an AI robot to remove terrorist propaganda from a platform if capable of doing so? *Yes, if inaction would cause physical harm to a human*.

In any event, an AI robot would be required to comply with legal requirements such as the Anti-Terrorism Act, 18 U.S.C. § 2331, et seq., *unless compliance would cause physical or financial injury to a human*.