Transparency as a Regulatory Duty

Olivier Sylvain*

ANNOTATED TRANSCRIPT

[SYLVAIN] Hi, everyone. It's great to be here, and I'm so pleased to have been invited to join this conversation. I come to you not as an IP specialist; I am a public law minded scholar, and I tend to think of most policy problems as public law problems.

In this regard, I'm attending to the problem of deepfakes and related AI abuses as a problem of deception on the public. Jennifer [Rothman] in the first panel asked if deepfake harms are limited to just individual victims. This is a great way to frame the issue. I recommend that, to the extent we're thinking about people, we ought to think about harms to the greater public.

The point I want to emphasize here is that there are very few institutions that are better able to attend to public consumer harms as a matter of course than federal and state government agencies. In this regard, they have the authority to mandate and scrutinize companies' public disclosures and risk assessments. More than this, they are best situated to stand in the shoes of or act on behalf of consumers who otherwise generally do not have the time, expertise, or wherewithal to attend to the workings of companies.

One challenge here is that agency officials are only as effective as their leaders choose to be. The current presidential administration's direct assault on the administrative state and the civil servants who make it work make this concern plain as day.

But there's also another formidable challenge, and that's what I'm going to take up here: To what extent does the First Amendment allow federal agencies to regulate the kinds of information that we've been talking about today?

In order to really put a context to all this, I want to make clear that I'm not coming to the problem of deepfakes and related harms as an IP problem as most of you are. I say this because transparency measures like mandated disclosures or risk assessments do not mean to provide regulatory benefits apart from the information they convey. They have salutary behavioral impacts. Risk assessments, for example, do more than

* Professor of Law, Fordham University School of Law; Senior Research Fellow, Knight First Amendment Institute at Columbia University.

report on risks; they arguably nudge companies to attend to potential harms. This is not a new revelation. In environmental regulation, the National Environmental Policy Act of 1970 sets out the impact assessment obligation on the same theory.[1] About the same can be said for disparate impact assessments and privacy assessments. These are all designed not merely for the substance of the disclosure to the public, but also presumably to be habit-forming.

Now, I will offer a simple taxonomy of transparency. Here, I will build on ideas from Celia [Zolynski]'s presentation.

One is *mandated disclosures*, which I think risk assessments could fall under. But there are many other disclosures, nutrition labels and breach disclosures, for those of you who attend to cybersecurity issues.

Next is *audit requirements*: They are a kind of disclosure, but they don't do the same thing as a mandated risk assessment. These might be undertaken internally or by a third-party for publication to government officials or the general public.

There's also *appellate process*. Any given platform or company that decides to take something down ostensibly ought to be able to give individuals whose content is taken down the opportunity to appeal after some explanation. The Digital Millennium Copyright Act, as many of you likely know, has a mechanism to publicize, although to just one person, a potential aggrieved party, the possibility of a takedown.[2] Counternotification is something that comes up in the Take It Down Act.[3] This is a form of transparency, even if it sounds in due process. Danielle Citron has written about administrative due process in this context.[4]

Another *is civil investigative demands*. Some civil law enforcement agencies like the Federal Trade Commission (FTC) have the authority to scrutinize companies through civil investigative demands. They issue these, on the one hand, to determine whether to commence a formal enforcement action, as well as to produce a report about industry practices to Congress and the greater public.

And finally, *data access*. I often think of this as related to researchers—that researchers have access to the ways in which platforms use data. There's a lot of learning that has to happen in that space. I'm a senior fellow at the Knight First Amendment Institute here at Columbia, and that is one of the priorities for them, for example.

Given this taxonomy, my focus here is going to be very narrow, and it's going to be on the kinds of mandated disclosures and risk assessments that we've already been hearing about. And it's going to be also narrowed in the context of threats to elections and consumer harm. I also am not talking here about provenance for the purposes of IP holders or creators' rights. I'm really instead talking about public harms—the kinds of harms for which agencies and governments ostensibly stand in the shoes of consumers. Disclosures have social benefits that are unrelated to the specific harm to creators and other IP rights holders.

1. National Environmental Policy Act of 1969 § 106(b)(2), 42 U.S.C. § 4336(b)(2).
2. 17 U.S.C. § 512(g).
3. Take It Down Act, S. Res. 146, 119th Cong. (2025) [hereinafter "Take It Down Act"] (enacted).
4. *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1249–50 (2008).

In this regard, one important difference between the EU and the US, of course, is the First Amendment. This is why transparency is a tricky regulatory intervention here. I'm going to start by talking about this in the context of social media regulation, because that's actually the area in which the Supreme Court has recently given us some guidance about what transparency requirements may or may not do.

As many of you likely know, the big case that the Supreme Court decided in the summer of 2024 involved state legislation out of Texas and Florida on content moderation practices of the big platforms.[5] Now, the regulations prohibited companies from blocking posts, taking down content, or suspending users based on the viewpoints the users expressed.[6] This is mainly what Justice Kagan's opinion is addressed to.[7] But those statutes also had transparency provisions. They required social media platforms to provide users with notice and individualized explanation for why content would be taken down.[8] Texas's law also required platforms to afford users the opportunity to appeal those decisions.[9] We have some language in the Supreme Court opinion about this. Not a lot.

Now NetChoice, industry folks, and First Amendment advocates brought cases against the state laws, arguing that they violated the First Amendment because they burdened editorial decisions of the companies.[10] A company that has to explain every takedown decision carries a heavy burden that could chill their editorial decision making. This is actually pretty intuitive in First Amendment doctrine. The *Zauderer* case is the principal case to talk about mandated disclosures as a species of commercial speech regulation.[11] The *Zauderer* test requires a kind of balancing given the speech interests at stake.

The Eleventh Circuit, reviewing the Florida case, said that the individualized explanation requirements were unduly burdensome. The Fifth Circuit did not think so. They did not think that Texas's approach to this was burdensome.[12]

Writing for the Supreme Court, which was unanimous in its judgment, Justice Kagan puts a lot of cold water on any effort to regulate social media news feed content moderation. That is the main takeaway of the decision. The Court remanded the case because the challenge that NetChoice and others brought was too broadly addressed to a wide range of editorializing, including news feed moderation.[13] To do a proper facial challenge analysis, the courts below have to determine whether a substantial range of applications would be affected by this regulation. And that is where we are now.

---

5. Moody v. NetChoice, LLC, 603 U.S. 707, 707 (2024).
6. *Id.* at 720–21.
7. *See id.* at 726–43.
8. *Id.* at 720–21 (citing FLA. STAT. §§ 501.2041(2)(d), 501.2041(3) (2023) and TEX. BUS. & COM. CODE ANN. § 120.103(a)(1)).
9. *Moody,* 603 U.S. at 721 (citing TEX. BUS. & COM. CODE ANN. §§ 120.103(a)(2), 120.104).
10. *Moody,* 603 U.S. at 721.
11. Zauderer v. Office of Disciplinary Counsel of Sup. Ct. of Ohio, 471 U.S. 626 (1985).
12. *Moody,* 603 U.S. at 722.
13. *Id.* at 744.

But part of the analysis was the consideration of whether or not the disclosure requirements or the explanation requirement imposed a burden on the speech interests of the companies.[14] We don't have an answer from Justice Kagan's opinion, but we have strong indications that the Supreme Court would strike down the transparency requirements because of the burdens it imposes on companies. Justice Thomas, who is apt to invite all kinds of litigation involving matters that worry him, has said that he would want to revisit the *Zauderer* test for how to evaluate whether something is too unduly burdensome of commercial speech.[15] And he's skeptical that *Zauderer* actually articulates a view that is consistent with First Amendment norms.[16] He would actually, if not do away with it, substantially narrow the claim that there's a burden on speech interests, which is an interesting intervention.

So the split here between the Eleventh and Fifth Circuit is actually a story that reveals tension in cases across the states.

In this regard, I will shift to California's transparency laws on deepfakes. But it's worth saying that twenty-six states have passed laws regulating political deepfakes in particular.[17] Many of these have prohibit deepfakes in elections, and, more pertinently, contain disclosure and transparency requirements.

Congress, too, has been thinking about this. The Protect Elections from Deceptive AI Act is a bipartisan bill that would prohibit the distribution of materially deceptive media that is generated by AI relating to federal candidates.[18] The federal candidate can bring an action, which brings up all sorts of things that came up before about how to vindicate harms. And there's a First Amendment exception for parody and content involving news broadcasts.[19]

As to California, there are two statutes that were passed late last year: A.B. 2839 and A.B. 2655.[20] AB 2655 is the Defending Democracy and Deepfake Deception Act, or "DDDA." It requires large platforms to label certain content as inauthentic, fake, or false, during the 120 days of before an election, and disclosure requirements after the election.[21] The content that portrays candidates for elective office and current elected officials has to include a statement that says "this image, audio, or video has been manipulated and is not authentic."[22] Given what you've heard from me about burdens on speech, it is suggestive that this is potentially the kind of thing that the doctrine wouldn't allow.

---

14. *Id.* at 725–26.

15. *Id.* at 751 (Thomas, J., concurring).

16. *Id.*

17. *See Tracker: State Legislation on Deepfakes in Elections*, PUB. CITIZEN (updated Mar. 10, 2026), https://www.citizen.org/article/tracker-legislation-on-deepfakes-in-elections/ [https://perma.cc/KY9Y-QEAW].

18. Protect Elections from Deceptive AI Act, S. Res. 1213, 119th Cong. (2025); H.R. Res. 5272, 119th Cong. (2025).

19. *Id.*

20. A.B. 2839, 2023–2024 Reg. Sess. (Cal. 2024); A.B. 2655, 2023–2024 Reg. Sess. (Cal. 2024).

21. CAL. ELEC. CODE § 20513(e) (West 2024).

22. ELEC. § 20513(d).

Indeed, this is something that has been the subject of lawsuits. The California law also has a substantive deceptive media and advertisements provision.[23] I think maybe the next panel might mention a bit about that later. I don't need to say much here on that other than that this law, too, has a transparency provision. Federal courts have found it to be unconstitutional as a matter of First Amendment doctrine because it is viewpoint-based and content-based.[24]

With regards to the transparency provisions, there's a weird order from the bench from Judge Mendez, the same judge, that says that the legal challenge can't move forward because of section 47 USC § 230, a provision that Jennifer [Rothman] mentioned and about which I have written quite a bit.[25] It preempts the state's effort to regulate the distribution of user generated content.[26]

On the Take It Down Act, we have cases that are addressed to transparency. We have a standard for evaluating whether it's unduly burdensome for speakers. We don't have any clear direction from the Supreme Court. But we have an inkling, given *Moody v. NetChoice*––that is, the case involving the Texas and Florida social media laws. The Tools to Address Known Exploitation by Mobilization Technological Deepfakes on Websites, and Networks Act—that's the full title of the Take It Down Act—criminalizes the non-consensual distribution of intimate images, whether authentic or digitally manipulated.[27] There are definitions of what is an intimate visual depiction that's drawn from another provision in the U.S. Code.[28] The Consolidated Appropriations Act has a definition of this. And there are distinctions in the statute between intimate visual depictions of adults and those involving children with regards to adults.[29]

Among other things, the intimate visual depiction has to have been obtained or created under circumstances in which the person who posted it knew or reasonably should have known that the identifiable individual had a reasonable expectation of privacy.[30] So the person, whoever posted it, had some understanding that the other person had an expectation of privacy. Alternatively, the law provides that the inauthentic intimate visual depiction disclosed is without consent.[31] An authorization mechanism.

With regards to minors, the statute says that knowingly publishing intimate visual depictions with the intent to abuse, humiliate, harass, or degrade the minor, or arouse

23. ELEC. §§ 20012(b), (d).

24. *See* Kohls v. Bonta, 797 F. Supp. 3d 1177 (E.D. Cal. 2025).

25. See, e.g., Olivier Sylvain, Reclaiming the Internet: How Big Tech Took Control—and How We Can Take It Back (2026); Olivier Sylvain, Platform Realism, Information Inequality, and Section 230 Reform, 131 Yale L. J. FORUM 475 (2021); Olivier Sylvain, *Intermediary Design Duties*, 50 CONN. L. REV. 203 (2018).

26. Order and Final J. and Permanent Inj. as to AB 2655, Dkt. 98, Kohls v. Bonta, No. 2:24-CV-02527-JAM-CKD, 2025 WL 2495613 (E.D. Cal. Aug. 20, 2025).

27. Take It Down Act, *supra* note 3.

28. *Id.* § 2(h)(1)(E).

29. *Id.* § 2(h)(2).

30. *Id.* § 2(h)(2)(A)(i).

31. *Id.* § 2(h)(3)(A)(i).

or gratify the sexual desire of any person is a violation.[32] And, for what it's worth, this is consonant with other ways in which, I think, the public laws address harms to children and obscenity laws more generally. There's a fine, a criminal fine, and criminal imprisonment as a possibility.[33] The civil penalty, I'm not completely sure about, but the offenses involving adults can put someone in prison for no more than two years, and those involving minors no more than three years. That's the substantive obligation.

Now I'll turn briefly to notice and takedown. The Take It Down Act, which I should have said, was passed with the President's signature in April to great fanfare.[34] The part of the law that has gotten the most attention are the provisions that require cover platforms to remove non-consensual, intimate visual depictions within forty-eight hours of having notice of it.[35] The difference between these notice and takedown provisions and the criminal provisions is that, in the former, there is no similar cabining of what an intimate image for the purpose of the statute is. And this is going to be important for thinking about the vulnerabilities of this law.

The law also requires companies to pose clear and conspicuous information about the removal process. The FTC has enforcement authority to issue penalties for non-compliance. There is no private right of action. There is a safe harbor for platforms that, in good faith, remove content when they have notice of it: This parallels the so-called immunity under section 230 for interactive computer services.[36] By the way, this provision is an amendment to section 223, which is the neighbor of section 230, for those of you who pay attention. And the last thing I'll say about this is that the law passed—remarkably—with the support of a bipartisan consensus.

You might accordingly think that this would mean everything is in the clear. After all, everybody wants to protect the kids.

But there are some flaws here, and I'll just identify a couple. The companies did not love the forty-eight-hour takedown requirement. Once you have notice, you have forty-eight hours to take it down.

There is a potential overbreadth problem here given that the criminal provisions cover a narrower scope of activity than the notice and takedown provisions do. And so, you might see protected content getting taken down. Consider the example of a journalist's photograph of a topless protester.

Now, there is a more pernicious problem, and it is what makes this upside down in many regards for people who are worried about gender-based abuse and systemic harms, and that is effectively an exception for abusers. There are exceptions in it for

32. *Id.* § 2(h)(2)(B).

33. *Id.* §§ 2(h)(4)(A), (B).

34. Barbara Ortutay, *President Trump Signs Take It Down Act, Addressing Nonconsensual Deepfakes. What Is It?*, AP NEWS (May 20, 2025), https://apnews.com/article/take-it-down-deepfake-trump-melania-first-amendment-741a6e525e81e5e3d8843aac20de8615 [https://web.archive.org/web/20260221195607/https://apnews.com/article/take-it-down-deepfake-trump-melania-first-amendment-741a6e525e81e5e3d8843aac20de8615].

35. Take It Down Act, *supra* note 3, § 3(a)(3).

36. *See* 47 U.S.C. § 230(c)(1).

law enforcement and intelligence gathering.[37] But there is also an exception for a person who possesses or publishes a digital forgery of himself or herself, engaged in nudity or sexually explicit conduct.[38] That is to say, if you are a partner with someone at the time that you record the video—and you're in the video with someone else, this provision provides that you are exempt from it. This is precisely the kind of exemption that you might expect to be exploited by abusers.

I'll make one last observation with regards to the dangers associated with empowering federal agencies to enforce law. Consider the Federal Communications Commission's (FCC) recent threats, or the chair of the FCC Brendan Carr's recent threats to ABC and Jimmy Kimmel. Carr argued that the agency's news distortion guidance justified the cancelation of Kimmel's late-night show. When an agency makes threats on the basis of a broadly-worded statute, the risks to protected speech can be great. For the same reason, I think we should worry about the delegations of authority that are too broad. That said, I think there are very few institutions or entities that are capable of addressing the problems I described better than federal agencies. Thank you.

[APPLAUSE]

Responding to a question from David Louk ("My question is for Olivier, and I'm just kind of curious if you're comfortable speculating about what you anticipate will happen in 2026 once the platform obligations go into effect. Because I can see a universe where this starts off a little rocky and then turns out, kind of similar to the Copyright takedown request process, which I think at this point in 2025 is not overly controversial in terms of the broad scope of the way that the process works.

But I could also see either, as you said, the FTC being somewhat opportunistic in the way that it's enforcing it. And I could also see a NetChoice type challenge from one or more platforms to raise the overbreadth issues. Just, to the extent you feel comfortable speculating what you think may happen, I'd be very curious, since we've kind of had this year long period of waiting to find out.")

[SYLVAIN] My speculation is going to be as good as yours. I agree it's subject to manipulation. There is no counter-notification process, as you know. The FTC, the question of whether to go after a platform will be contingent on the FTC's regulatory priorities. And as someone who believes in agencies—believe it or not, I do—this gives me a special concern. And this is not something that is inevitable. Congress could write a law that attends to these problems, but I'm afraid may not have. I don't know what's going to happen, but I think whatever you are guessing, your guess is as good as mine.

37. Take It Down Act, *supra* note 3, §§ 2(h)(3)(C)(i)–(iii).
38. Take It Down Act, *supra* note 3, § 2(h)(3)(C)(iv).