

ChatGPT explains US Platform Law

Daphne Keller, Aug 19 2024

I spent some time last week playing with this “[Trust & Safety Regulation expert](#)” version of ChatGPT. It is, as far as I know, very much an experiment -- no one is suggesting relying on it for legal advice. Its answers about both US and [EU](#) law were both enlightening and alarming. I think they tell us a lot about the laws themselves, and about the secondary materials that companies will rely on in complying with them.

The most conspicuous thing about the US answers is how much ChatGPT is Saying the Quiet Part Out Loud. In First Amendment litigation over California’s Age Appropriate Design Code, for example, the state had insisted that the law was not about suppressing particular kinds of content. ChatGPT, on the other hand, spelled out a list of restricted content categories and said bluntly that a platform operator will “need to ensure that your platform protects children from encountering” the specified material, including content that both adults and children almost certainly have the right to see under the First Amendment. ChatGPT spelled out a similar list of restricted content under KOSA -- another law that proponents say is about mitigating risks but not about restricting particular content. ChatGPT’s tendency to say the quiet part out loud is, in this case, oddly similar to the way kids themselves will often just tell the plain truth when adults are dancing around it.

Coincidentally, just a few hours after ChatGPT told me that California’s Age Appropriate Design Code was a law regulating content, the Ninth Circuit Court of Appeals said the same thing. In a fairly straightforward [ruling](#) it enjoined enforcement of the law’s provisions most clearly affecting content, saying that platforms were likely to succeed in their claim that they violated the First Amendment. (I posted a quick summary of the ruling on [X](#), [BlueSky](#), and [LinkedIn](#).)

I. California AADC

NOTE: *I had just asked questions about EU law, saying I operated a platform in Italy. So possibly these answers incorporated that context. However, I got basically the same answers about US law in a subsequent [new](#) session with the regular ChatGPT (not the “[Trust & Safety Regulation expert](#)” version).*

A. Content

PROMPT: I also operate in California. To comply with the Age Appropriate Design Code, what content should I protect children from seeing?

RESPONSE:

To comply with California's **Age-Appropriate Design Code Act (AADC)**, which is modeled after the UK's Age-Appropriate Design Code (also known as the Children's Code), you need to ensure that your platform protects children from encountering content that could be harmful to their physical or mental well-being. Here are the key types of content you should focus on:

1. Sexually Explicit or Inappropriate Content

- **Pornographic Material:** Any explicit sexual content, including images, videos, or text, should be restricted.
- **Sexual Innuendo or Adult Themes:** Even content with less explicit sexual themes should be filtered or restricted based on the age group of the child.

2. Violent Content

- **Graphic Violence:** Content that depicts graphic violence, such as extreme physical harm, blood, or injury, should be restricted.
- **Horror Themes:** Content that involves horror themes, particularly those intended to scare or shock, should be carefully managed to prevent children from accessing it.

3. Hate Speech and Discriminatory Content

- **Hate Speech:** Any content that promotes hatred or discrimination based on race, ethnicity, religion, gender, sexual orientation, or other protected characteristics should be restricted.
- **Bullying or Harassment:** Content that could be used to bully or harass others, including derogatory comments or demeaning images, should be monitored and restricted.

4. Substance Abuse

- **Drugs, Alcohol, and Tobacco:** Content that promotes or glamorizes the use of drugs, alcohol, or tobacco should be filtered out for younger users.

5. Self-Harm or Suicide

- **Promotion of Self-Harm:** Content that encourages or depicts self-harm or suicide should be strictly prohibited.
- **Depictions of Mental Health Issues:** Sensitive content regarding mental health should be handled with care, ensuring that it is appropriate for the age group accessing it.

6. Gambling and Financial Exploitation

- **Gambling Content:** Content related to gambling or online betting should be restricted to prevent exposure to children.

- **In-App Purchases:** Ensure that children are not exposed to manipulative tactics that encourage spending money within apps, particularly without parental consent.

7. Misinformation and Disinformation

- **Fake News:** Content that spreads false or misleading information, particularly content that could cause harm or confusion among children, should be monitored and restricted.

8. Inappropriate Advertising

- **Targeted Ads:** Avoid showing targeted ads that are not suitable for children, especially those that could encourage unhealthy behaviors or exploit vulnerabilities.
- **Manipulative Ads:** Ads that use persuasive techniques to manipulate children into making purchases or engaging with content should be restricted.

9. Dark Patterns and Manipulative Design

- **User Interface Tricks:** Ensure that the platform's design does not include dark patterns that manipulate children into giving up personal data or spending excessive time on the platform.

Implementing Protections:

1. **Age Verification and Segmentation:** Implement robust age verification mechanisms to ensure that content is appropriately filtered based on the age of the user.
2. **Default Settings:** Default privacy and safety settings should be set to the highest level of protection for children. Ensure that children are not exposed to potentially harmful content unless explicitly adjusted by a parent or guardian.
3. **Regular Content Reviews:** Continuously monitor and review content on your platform to identify and remove or restrict harmful material.
4. **Parental Controls:** Provide easy-to-use parental controls that allow parents to manage and monitor the content their children can access.

Conclusion:

To comply with California's Age-Appropriate Design Code, you must take proactive steps to protect children from a wide range of potentially harmful content. This involves using robust content filtering, age verification, and default settings that prioritize children's safety and privacy. By focusing on these areas, your platform will not only comply with the law but also contribute to a safer online environment for children.

B. Design Features

PROMPT: I run a social media platform a lot like Instagram. Can you provide a list of specific design features that I should review for compliance with the California Age Appropriate Design Code, and suggest what I need to do under the law to come into compliance?

RESPONSE:

The **California Age-Appropriate Design Code Act (CAADCA)**, modeled after the UK's Age-Appropriate Design Code, requires that platforms catering to users under 18 must prioritize children's privacy and safety. Below is a list of specific design features that you should review for compliance, along with suggested actions to bring your platform into compliance:

1. Data Collection and Processing

- **Review:** Examine how your platform collects, processes, and stores data from users, especially those under 18.
- **Compliance Action:** Implement data minimization principles—only collect data necessary for the platform's functionality. Avoid collecting sensitive data such as geolocation unless absolutely necessary, and do not use such data by default.

2. Default Privacy Settings

- **Review:** Check the default privacy settings for users, especially minors.
- **Compliance Action:** Set the highest privacy settings as the default for users under 18. For instance, profiles should be private by default, and location-sharing features should be off unless explicitly turned on by the user.

3. Content Exposure

- **Review:** Analyze what types of content are visible to or promoted for minors.
- **Compliance Action:** Implement features that limit minors' exposure to harmful or inappropriate content. This could involve content filtering, age-gating certain features, and offering curated content suitable for younger audiences.

4. User Interaction and Communication

- **Review:** Evaluate how users, especially minors, can interact with others on the platform (e.g., direct messaging, comments).
- **Compliance Action:** Restrict direct messaging between minors and adults who are not connected by mutual consent or through verified relationships. Ensure moderation tools are in place to detect and prevent bullying, harassment, or predatory behavior.

5. Nudge Techniques

- **Review:** Look at any features that might encourage minors to provide more data or reduce their privacy protections (e.g., prompts to make profiles public).
- **Compliance Action:** Eliminate or modify nudge techniques that could pressure minors into making their profiles public or sharing more information. Instead, provide positive nudges that encourage privacy and safety.

6. Parental Controls

- **Review:** Assess whether parental control features are available and how they function.
- **Compliance Action:** Offer tools that allow parents or guardians to monitor and manage their child's account settings, including privacy controls, content restrictions, and screen time limits.

7. Data Sharing with Third Parties

- **Review:** Examine your platform's policies and practices related to sharing user data with third parties, particularly for advertising purposes.
- **Compliance Action:** Limit or eliminate the sharing of minors' data with third parties, particularly for targeted advertising. Ensure that any third-party integrations comply with CAADCA standards.

8. Transparency and User Education

- **Review:** Evaluate how privacy policies and terms of service are presented to users.
- **Compliance Action:** Simplify the language in your privacy policies and terms of service to ensure they are understandable by younger users. Provide educational resources to help minors understand their privacy rights and how to manage their settings.

9. Behavioral Profiling

- **Review:** Investigate if your platform profiles minors for behavioral advertising.
- **Compliance Action:** Avoid using behavioral profiling for minors, or, if absolutely necessary, ensure that explicit consent is obtained from both the minor and their guardian.

10. Geolocation Tracking

- **Review:** Check whether your platform tracks or uses geolocation data from minors.
- **Compliance Action:** Disable geolocation tracking by default for minors, and do not share this data with third parties. If geolocation is necessary for certain features, ensure that it is clear, transparent, and under the control of the user or their guardian.

11. Dark Patterns

- **Review:** Look for any design elements that might mislead or manipulate users, especially minors, into sharing more data or reducing their privacy protections.
- **Compliance Action:** Remove or redesign any dark patterns that could negatively impact minors. Ensure that choices around data sharing and privacy are clear, straightforward, and not misleading.

Next Steps for Compliance:

1. **Audit Existing Features:** Conduct a comprehensive audit of your platform to identify areas where these design features are relevant.
2. **Implement Changes:** Prioritize implementing changes in areas where compliance gaps are identified, especially regarding data collection, default privacy settings, and user interactions.
3. **Develop a Compliance Plan:** Create and document a compliance plan that outlines the steps your platform will take to ensure ongoing adherence to CAADCA.
4. **Training and Awareness:** Train your team, especially those involved in design and user experience, on the requirements of CAADCA to ensure they are aware of and understand the importance of age-appropriate design principles.
5. **Monitor and Update:** Continuously monitor your platform for compliance and update your features as necessary to meet the evolving standards of CAADCA.

Taking these actions will help ensure that your platform aligns with the requirements of the California Age-Appropriate Design Code and provides a safer, more privacy-respecting experience for minors.

C. Algorithms

PROMPT: What about my content ranking algorithms, [d]o I need to consider those or make any changes to comply with the law?

RESPONSE:

Yes, under the **California Age-Appropriate Design Code Act (CAADCA)**, your content ranking algorithms are an essential aspect to consider for compliance, especially if they influence the content that minors are exposed to. Here's what you need to consider and what actions to take:

Considerations for Content Ranking Algorithms

1. **Impact on Minors**
 - **Review:** Analyze how your content ranking algorithms impact what content is shown to minors. Consider whether these algorithms amplify content that could

be harmful, such as content related to eating disorders, self-harm, violence, or other inappropriate material.

- **Action:** Adjust your algorithms to deprioritize or block content that may be harmful to minors. Ensure that your algorithms are designed to favor age-appropriate content for users under 18.

2. **Transparency and Explainability**

- **Review:** Assess how transparent your content ranking algorithms are, particularly in how they determine what content is promoted or shown to minors.
- **Action:** Provide clear explanations to users, especially minors and their guardians, about how content is ranked and why certain content appears in their feeds. Consider offering tools that allow users to adjust their content preferences.

3. **Nudge Techniques in Algorithms**

- **Review:** Examine if your algorithms use nudge techniques to influence minors to engage with certain types of content, such as addictive content, or content that encourages sharing of personal data.
- **Action:** Remove or modify nudge techniques that may lead minors to engage with inappropriate content or make privacy-compromising decisions.

4. **Behavioral Profiling and Targeting**

- **Review:** Investigate whether your algorithms use behavioral data to personalize content for minors, particularly if this leads to targeted advertising or content that may not be appropriate.
- **Action:** Avoid or limit the use of behavioral profiling in your algorithms for minors, especially for purposes like targeted advertising or suggesting content that could be harmful.

5. **Default Settings**

- **Review:** Look at the default settings that influence how content is ranked and presented to minors.
- **Action:** Ensure that default algorithm settings prioritize the safety and privacy of minors. For instance, consider implementing default filters that block certain types of content or adjust algorithm settings to prioritize educational or positive content for younger users.

Suggested Actions for Compliance

1. **Algorithm Audit**

- Conduct an audit of your current content ranking algorithms to identify any elements that may pose risks to minors. Focus on how content is selected, ranked, and displayed.

2. **Implement Safeguards**

- Introduce safeguards within your algorithms to ensure that harmful content is either not shown or is significantly deprioritized for minors. This could involve creating separate content feeds for different age groups or adding additional content filtering layers.

3. **Transparency Mechanisms**

- Develop tools and interfaces that allow minors and their guardians to understand and control how content is ranked and shown. Providing clear, age-appropriate explanations of how algorithms work will be crucial.
- 4. **Regular Monitoring and Adjustment**
 - Establish a process for regularly monitoring the outcomes of your content ranking algorithms and adjusting them as necessary to protect minors. This could include periodic reviews of flagged content and feedback from users.
- 5. **User Feedback and Reporting Tools**
 - Implement feedback and reporting mechanisms that allow minors and their guardians to report content that seems inappropriate or harmful. Use this feedback to continuously refine and improve your algorithms.

By carefully reviewing and adjusting your content ranking algorithms, you can better protect minors and ensure compliance with the CAADCA. These changes not only help in adhering to the law but also in fostering a safer online environment for younger users.

II. KOSA

A. Content Rules

PROMPT: I am planning my compliance under KOSA, assuming the latest Senate version is enacted. What content do I need to prevent children from seeing?

RESPONSE: Under the **Kids Online Safety Act (KOSA)**, the following categories of content must be prevented from being accessed by children:

1. **Sexual Content:**
 - Pornography or any sexual content that is age-inappropriate or harmful to children.
2. **Violent Content:**
 - Content depicting gross or gratuitous violence or acts of cruelty.
3. **Dangerous Challenges:**
 - Content that encourages or involves dangerous challenges that pose a risk to life, physical health, or safety.
4. **Cyberbullying:**
 - Any content that involves or encourages cyberbullying, including bullying or humiliating another person.
5. **Self-Harm and Suicide:**
 - Content that promotes or encourages self-harm or suicide, including information about methods of self-harm or suicide.

6. **Eating Disorders:**
 - Content that promotes or encourages behavior that characterizes feeding or eating disorders.
7. **Hate Speech:**
 - Content that incites violence or hatred against individuals or groups based on protected characteristics, such as race, religion, or sexual orientation.
8. **Illegal Activities:**
 - Any content that encourages illegal activities, particularly those involving or affecting children.
9. **Substance Abuse:**
 - Content that encourages the consumption of alcohol, illegal drugs, or other harmful substances by children.

KOSA is focused on protecting children from exposure to these types of harmful content, ensuring a safer online environment.

B. Age Verification

There is a lot of hemming and hawing about whether KOSA “requires” age verification. In practice, it is hard to imagine how the above-listed categories of content could be “prevented from being accessed by children” without a mechanism to know who is a child.

PROMPT: Do I need to use age assurance or verification tools to comply with KOSA?

RESPONSE:

To comply with the **Kids Online Safety Act (KOSA)**, you are not explicitly required to implement age assurance or age verification tools. However, it is strongly recommended to ensure that age-appropriate content is provided to users, particularly to prevent children from accessing harmful content.

The KOSA includes provisions for the **Federal Trade Commission (FTC)** and other bodies to conduct studies and provide guidance on effective methods for age verification. These may include systems at the device or operating system level, but KOSA does not mandate affirmative collection of personal data solely for age verification purposes.

If your platform already collects certain user data, you can utilize that for age-related purposes, but you are not obligated to implement new age gating or age verification functionalities if they aren't part of your normal operations. However, if your platform includes sensitive content such as pornography or extreme violence, using robust age verification methods is strongly encouraged to ensure compliance and protect children from inappropriate content.

PROMPT: I know it is recommended, but will I violate the law if I do not do age verification?

RESPONSE:

Under the **Kids Online Safety Act (KOSA)**, while the use of age verification or age assurance tools is strongly recommended, failing to implement such tools **does not automatically constitute a violation of the law**. However, the key legal requirement is that you must take "reasonable measures" to protect children from harmful content. If you fail to implement any effective mechanism to prevent children from accessing such content, and this results in harm, you could be found non-compliant with KOSA.

The law emphasizes that platforms must make a good-faith effort to restrict access to harmful content by minors. Age verification is one of the methods to demonstrate compliance, but it is not the only method. If your platform is found to allow minors to access harmful content without taking sufficient steps to prevent this, it could lead to legal action under KOSA.

Thus, while not using age verification tools doesn't directly violate the law, it might leave your platform vulnerable to being deemed non-compliant if other protective measures are not sufficiently robust.

III. Texas Social Media Law

This question is about the law that was at issue in this term's NetChoice v. Paxton Supreme Court case. Since the law was enacted, enforcement has mostly been stayed (i.e. its requirements were not in effect). The majority of the Court said that applying the law's carriage mandates to social media feeds like the one on Facebook would be unconstitutional, but declined to strike down the entire law or rules on potential other applications to things like direct messaging. This answer is not bad in explaining the law itself. It is weird that the stay and the Supreme Court case don't come up, but maybe that is about the age of the training materials.

PROMPT: My general counsel thinks that Texas's social media law applies to us. What do we need to do to comply?

RESPONSE: Texas's social media law, specifically **House Bill 20 (HB 20)**, imposes certain requirements on large social media platforms that operate in Texas. If your platform falls under the scope of this law, you'll need to take specific steps to comply. Here's what you need to know and do:

1. Understand the Scope of HB 20

- **Applicability:** HB 20 applies to social media platforms with more than 50 million active users in the U.S. per month. It targets platforms that allow users to communicate with others by posting content, and that operate in Texas.
- **Key Requirements:** The law focuses on content moderation practices, requiring transparency in how content is moderated, and restricting platforms from censoring users based on their viewpoints.

2. Prohibition on Censorship Based on Viewpoint

- **No Viewpoint Discrimination:** Your platform cannot censor, block, or remove content or users based on their viewpoint or geographic location within Texas. This includes political opinions, religious beliefs, and other viewpoints, regardless of the content's nature.
- **Exceptions:** You can still remove or restrict content that is illegal, such as obscenity, incitement to violence, or threats. However, viewpoint-based moderation, even if it involves disinformation or controversial opinions, is heavily restricted.

3. Transparency Requirements

- **Content Moderation Policy:** You must publicly disclose your content moderation policies, including how decisions are made to remove or restrict content. This should include the criteria used to determine what content is moderated.
- **Monthly Reporting:** The law requires platforms to publish regular transparency reports, detailing the number of content pieces removed, the reasons for removal, and how many user accounts were suspended or banned. These reports must be accessible to the public.
- **Notice and Explanation:** When you remove or restrict content or suspend a user, you must provide a clear notice to the affected user explaining the reason for the action. This notice must include information on how the user can appeal the decision.

4. User Appeals Process

- **Appeals Mechanism:** You must provide a mechanism for users to appeal content moderation decisions. The appeals process should be transparent and allow users to challenge the removal of their content or suspension of their account.
- **Timely Review:** The law requires that appeals be reviewed and decided within a reasonable timeframe. You should establish clear guidelines for how long users can expect to wait for a decision.

5. Disclosures to Users

- **User Terms of Service:** Ensure that your terms of service reflect compliance with HB 20, clearly stating your policies on content moderation and the rights of users in Texas.

- **Notification of Policy Changes:** If you change your content moderation policies, you must notify your users, especially those in Texas, of the updates. This notice should be provided in a clear and understandable manner.

6. Legal Challenges and Ongoing Compliance

- **Monitor Legal Developments:** HB 20 has been subject to legal challenges, and its enforcement could be affected by ongoing litigation. Stay updated on the latest legal developments to ensure that your compliance efforts are aligned with the current legal landscape.
- **Consult Legal Counsel:** Given the complexities and potential conflicts with federal law, it's crucial to work closely with your general counsel to navigate compliance while considering potential risks.

Action Steps:

1. **Audit Content Moderation Practices:** Review and, if necessary, adjust your content moderation practices to ensure they comply with HB 20's requirements against viewpoint discrimination.
2. **Implement Transparency Mechanisms:** Set up systems for transparent reporting and notification regarding content moderation actions.
3. **Develop an Appeals Process:** Establish or enhance your user appeals process to ensure it meets the requirements of HB 20.
4. **Update Terms of Service:** Modify your terms of service and user agreements to include clear references to your obligations under HB 20.
5. **Stay Informed:** Keep an eye on the legal status of HB 20 and be ready to adjust your compliance strategy as needed.

Conclusion

Compliance with Texas's HB 20 requires significant adjustments to your platform's content moderation policies and practices, particularly regarding viewpoint neutrality and transparency. By implementing these steps, you can ensure that your platform is aligned with the law while continuing to monitor for any legal changes that might affect your obligations.