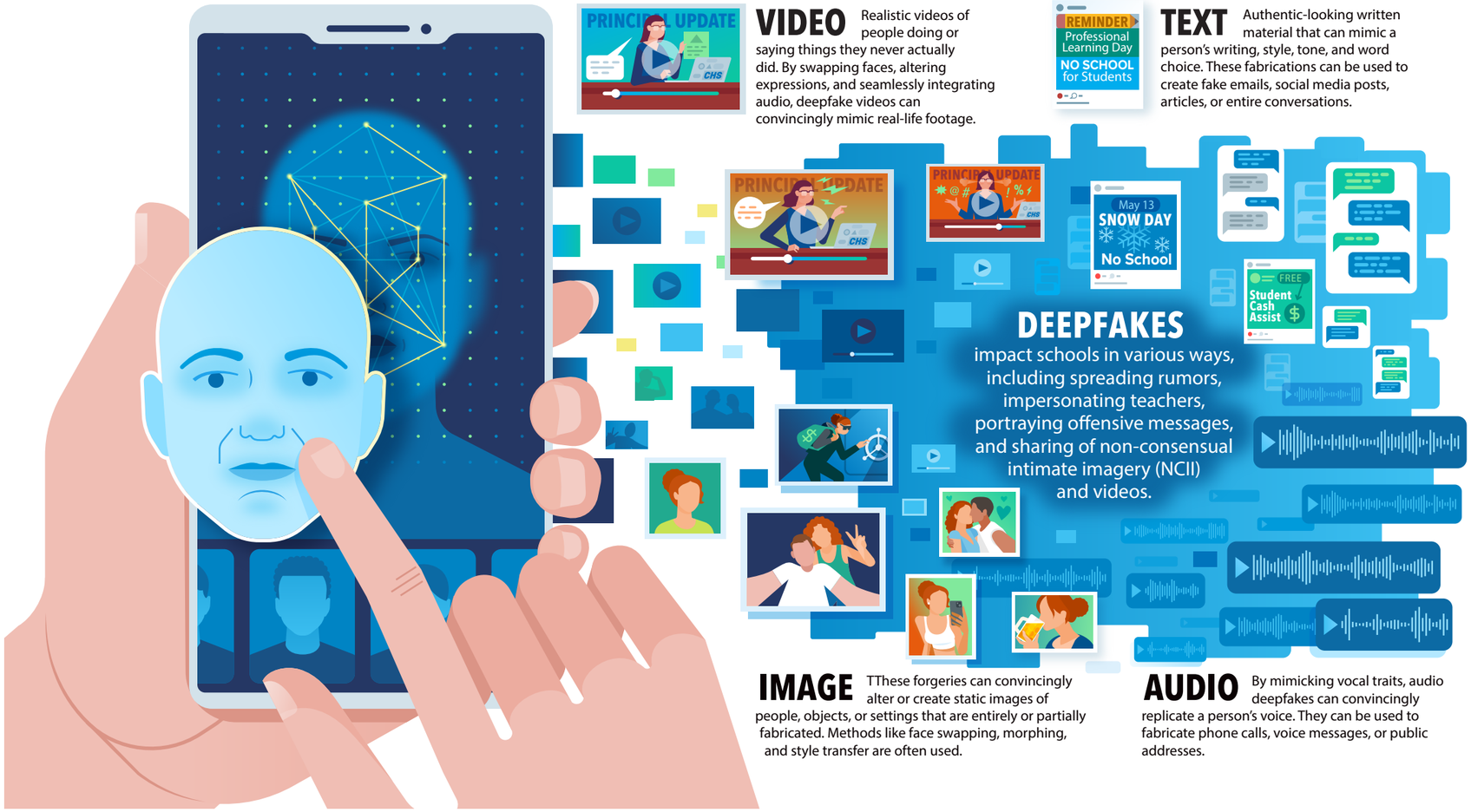
Deepfakes in School: Risks and Readiness

Deepfakes are synthetic images, videos, audio, and text, that are created or manipulated using AI and modeled after real people. Deepfakes can seamlessly alter or re-create the appearance and voice of individuals, making it appear as though they are saying or doing something they never actually did. This technology poses significant challenges for schools, as it can be used to target students and staff, spread disinformation, perpetrate fraud, and undermine trust. Deepfakes impact schools and school communities in various ways. They may be used to instigate or perpetuate rumors, embarrass individuals, or promote offensive messages. In particularly egregious circumstances, deepfakes may constitute non-consensual intimate imagery (NCII) and videos.



VIDEO Realistic videos of people doing or saying things they never actually did. By swapping faces, altering expressions, and seamlessly integrating audio, deepfake videos can convincingly mimic real-life footage.

TEXT Authentic-looking written material that can mimic a person's writing, style, tone, and word choice. These fabrications can be used to create fake emails, social media posts, articles, or entire conversations.

IMAGE TThese forgeries can convincingly alter or create static images of people, objects, or settings that are entirely or partially fabricated. Methods like face swapping, morphing, and style transfer are often used.

AUDIO By mimicking vocal traits, audio deepfakes can convincingly replicate a person's voice. They can be used to fabricate phone calls, voice messages, or public addresses.

Real-World Example

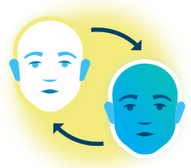
Deepfake non-consensual intimate imagery (NCII) can be generated by face-swapping, replacing one person's face with another's face, or digitally "undressing" a clothed image to appear nude. In the case where NCII involves minors, it may also be considered Child Sexual Abuse Material (CSAM). These deepfakes raise many of the same issues as non-synthetic NCII and CSAM, though potential offenders may not appreciate the serious, criminal implications. While many of these deepfakes may be created and shared outside of school, schools are required to address off-campus behavior that creates a "hostile environment" in the school. Consider how your school would respond to the below incident as it unfolds.

A student reports that they received a sexually explicit photo of a classmate and that the photo is circulating among a group of students.

How can your school leverage internal investigative tools or processes used for other technology or harassment incidents?

Administrators begin questioning students to determine the extent of the incident.

What process does your school use to reduce distribution, ensure the privacy of all students involved in an investigation, and provide appropriate support for the impacted individual?



The investigation reveals there is potential that the image is a deepfake.

How might the potential of a deepfake impact the investigation and response?

The administrator reviews existing policies to ensure integrity of the investigation and determines the extent of legal counsel and law enforcement involvement.

What policies and procedures does your school have that may apply?

Community dynamics are considered when constructing any public communication regarding the incident; all communication is consistent and mindful of privacy impacts.

What processes does your school have to ensure the privacy of students and minimize harm when communicating?

How Are Deepfakes Harmful?

BULLYING & HARASSMENT

MISINFORMATION & DISINFORMATION

PERSONAL SAFETY & PRIVACY

IMPACT ON THE LEARNING ENVIRONMENT

Readiness Checklist

Deepfakes can be created using commonly available online tools or proprietary programs and are increasingly sophisticated and hard to detect. While some jurisdictions have started to create requirements for transparency or authentication for all types of synthetic content, including deepfakes, those requirements do not yet have substantial reach. School leaders must be vigilant in addressing its potential impacts by identifying and mitigating potential harms while navigating this evolving challenge. Recent incidents have highlighted the need to protect students, staff, and administrators while addressing any incident. Administrators can better respond and mitigate potential harms by leveraging existing policies and procedures and considering:

- ☐ Educate and train students, staff, and parents about the impacts and consequences of deepfakes
- ☐ Keep current with laws and regulations and understand how laws apply to sharing student information, even when that information may be AI-generated
- ☐ When investigating incidents, consider that any digital media could be a deepfake and reliable detection is difficult
- ☐ Engage in open discussion on how to address potential incidents
- ☐ Establish communication protocols around what to communicate and to whom
- ☐ Provide support to the impacted individuals, and consider confidentiality and privacy of all parties when investigating and communicating about an incident
- ☐ Be aware of personal and community biases and norms
- ☐ Determine how existing policies and practices might apply, including policies on bullying, harassment, Title IX, sexting, technology use, disruption of school, misconduct outside of school, and impersonation of others on social media
- ☐ Update current policies and procedures to ensure they address deepfakes, including image-based sexual abuse and harassment
- ☐ Consider implementing policies for third-party vendors on how they should address a deepfake incident
- ☐ Evaluate if a sexually explicit deepfake incident qualifies as a form of sexual harassment
- ☐ Consult with legal counsel
- ☐ Work with local law enforcement to establish incident thresholds and response responsibilities
- ☐ Establish an after-action review process

For more resources visit **studentprivacycompass.org/deepfakes**

FUTURE OF PRIVACY FORUM | AI