Deepfakes and Synthetic Media

A Plain-Language Safety Guide
for Schools and Families

AstraEthica
astraethica.ai

2026 · Version 1.2

What This Guide Is (and Isn't)

This guide is not a technology manual. It will not teach you to build a deepfake, run a detection algorithm, or configure parental controls. It was written for any adult responsible for young people — teachers, counselors, school leaders, parents, caregivers, coaches, mentors, faith leaders, and youth workers — who wants to understand a risk that is growing faster than most institutions have prepared for.

The focus is on judgment, not forensics. Deepfake detection tools exist, but they are unreliable, expensive, and already falling behind the technology they chase. What lasts is a young person's ability to question what they see, talk to someone they trust, and respond wisely when something feels wrong.

Our goal is to lower panic and raise clarity. Most adults have heard the word "deepfake." Many picture a world leader giving a fake speech or a celebrity placed in a video they never made. Those things are real. But the version of this problem that most often reaches young people is quieter, closer to home, and more personal than a political hoax — and it is happening in schools right now.

You do not need to be a technology expert. You need to be steady, informed, and willing to have uncomfortable conversations.

> *"Video, images, and audio used to feel like proof. Now they can be edited or fabricated so convincingly that our eyes and ears are not enough."*

If a young person is in immediate danger, contact emergency services first.

Crisis Text Line: Text HOME to 741741. 988 Suicide and Crisis Lifeline: Call or text 988. NCMEC CyberTipline: 1-800-843-5678.

Short on Time?

If you read one section, read the Four Foundations. They are the heart of everything else in this guide.

If you are worried about something happening right now, start with Signals Worth Paying Attention To and If Something Goes Wrong.

If you work in a school, read When a Normal-Looking Clip Quietly Goes Wrong for scenarios you can use in staff conversations.

If you only have two minutes, turn to the last page: What Every Adult Should Know About Deepfakes and Young People.

What Deepfakes Actually Are — and Why They Matter Now

A deepfake is any video, image, or audio that has been created or altered by artificial intelligence to make someone appear to say or do something they never did. The word covers a wide range — a face swapped onto another body, a voice cloned from a short recording, a photograph manipulated to show something that never happened.

Three things have changed recently that make this an urgent concern for anyone working with young people.

The tools are now free and easy. A convincing fake nude image can be generated from a single ordinary photograph in less than half an hour, at no cost, by anyone with a phone or a web browser. Voice cloning requires as little as a few seconds of someone's voice — the kind of clip a teenager posts to social media every day. Dozens of apps designed specifically to strip clothing from photos are available through ordinary search engines, and some face-swap apps are rated suitable for young children in app stores.

The scale is enormous. The volume of synthetic media shared online has grown dramatically in the past two years and continues to accelerate. Reports of AI-generated child sexual abuse material have surged, with some organizations tracking increases of several hundred percent in a single year. In schools, surveys suggest that a significant number of high school students are

aware of a deepfake involving someone from their school, and a meaningful minority say a classmate has used AI to create explicit images of another student.

Detection is no longer reliable. In controlled studies, people struggle to identify high-quality deepfake video with any consistency. The old tells — extra fingers, blurry hairlines, unnatural blinking — have largely been corrected by newer tools. No automated detection system has been shown to work reliably in real-world conditions. The expert consensus is clear: we cannot detect our way out of this problem.

Four Foundations

These are lenses, not rules. Each describes a shift that can happen gradually, without anyone noticing, whenever synthetic media enters a young person's world.

FOUNDATION 1

Trust

Seeing used to mean believing. A photograph, a voice recording, a video — these were evidence. When a classmate showed you a clip, you trusted your own eyes. That assumption is no longer safe, and most young people have not yet adjusted.

The danger is not only that young people might believe something fake. It is also that they may stop believing things that are real. When any piece of media can be fabricated, authentic evidence becomes deniable. A student who is bullied can be told the proof is "just AI." A young person who witnesses something can doubt their own experience. This erosion of shared truth is one of the quietest and most serious harms of synthetic media.

Why this matters. Young people need to understand that trust in media now requires a step that used to be unnecessary: verification. Not suspicion of everything, but a habit of pausing before reacting — especially when something is shocking, upsetting, or too good to be true.

What healthy trust looks like. A young person who sees a surprising image or video and says, "Let me check before I share this," is practicing the right instinct. The goal is not cynicism. It is calm, confident skepticism.

FOUNDATION 2

Evidence

If someone cannot spot a deepfake by looking at it, what are they supposed to do? The answer is not a technical tool. It is a set of habits: asking where something came from, checking whether other sources confirm it, and looking for the original context before accepting a single version of events.

These are the same skills that good journalists and researchers use. They are not complicated. But they are rarely taught — and when media looks real, most people skip them entirely.

Why this matters. When a young person encounters a piece of media that claims to show something alarming, their first reaction is usually emotional. Verification skills give them a second step — not a replacement for emotion, but a way to act on it wisely.

What good evidence habits look like. Before sharing, ask: Where did this come from? Who posted it first? Do other sources confirm it? Is there a reason someone would want me to believe this? These questions do not require technology. They require patience.

FOUNDATION 3

Harm

The most common way deepfakes hurt young people is not through political disinformation. It is through intimate image abuse. In school after school, the pattern is the same: a student takes an ordinary photograph of a classmate — from social media, a school event, a group chat — and runs it through an app that generates a fake explicit image. The image is shared. The damage is immediate and lasting.

Girls are overwhelmingly the targets. The vast majority of deepfake intimate imagery involves female subjects. Boys are more often the creators — and in many cases, they describe what they did as a joke or a dare, not understanding the legal and emotional consequences until it is too late.

But boys are harmed too, differently. Financial sextortion — where a scammer uses real or fake intimate images to demand money — disproportionately targets teenage boys. The shame and fear involved have, in documented cases, led to devastating outcomes.

Why this matters. Young people need to understand that creating, possessing, or sharing a fake intimate image of someone is a form of abuse. In most places it is also a crime. The fact that the image is not “real” does not make the harm less real.

What awareness of harm looks like. A young person who hears about a deepfake of a classmate and refuses to look at it, share it, or joke about it is demonstrating the response that matters most.

FOUNDATION 4

Power

Deepfake technology gives one person the ability to control another person’s image, voice, and reputation without their knowledge or consent. That is a power imbalance, and it tends to follow

the same lines that power imbalances always follow: older over younger, popular over isolated, male over female.

Understanding this dynamic helps young people see synthetic media not just as a technology issue but as a fairness issue. Who gets targeted? Who does the targeting? Who has the resources to fight back, and who does not?

Why this matters. When adults frame deepfakes only as a tech problem, they miss the human patterns underneath. A young person who understands the power dimension is better equipped to recognize abuse, support a peer, and challenge the behavior — not just detect the technology.

What an understanding of power looks like. A young person who asks "Who benefits from this being believed?" or "Who is being hurt by this?" is thinking about power — even if they have never used that word.

> *"The skills that protect young people are not technical — they are relational: trust, evidence, awareness of harm, and an understanding of power."*

When a Normal-Looking Clip Quietly Goes Wrong

These composites describe patterns, not specific young people. Each blends details from real situations that schools and families have encountered.

The Forwarded Image

What adults might see. A group of students is buzzing about something on their phones. A teacher hears laughter, then whispers, then silence when an adult approaches.

What may actually be happening. A fake intimate image of a classmate has been generated and is circulating through a group chat. Most students who see it know it is wrong but feel trapped: sharing it makes them complicit, but saying nothing feels like the only way to stay out of trouble. The student in the image may not know yet.

What might open the conversation. “If you saw something online that you knew was hurting someone, what would make it easier to tell an adult?”

The Perfect Prank

What adults might see. A funny video of a teacher or principal surfaces online. Students find it hilarious.

What may actually be happening. The video was generated using AI, placing the staff member’s face and voice into an embarrassing or offensive scenario. The student who made it sees it as a joke. The staff member may face professional consequences, emotional distress, or community backlash before anyone confirms the video is fake.

What might open the conversation. “Even if something is clearly a joke to you, what happens to the person in the video who did not agree to it?”

The Proof That Was Not Proof

What adults might see. A student shows a recording as evidence that another student said something cruel. The recording sounds authentic.

What may actually be happening. The audio was generated using a voice-cloning tool. The accused student genuinely did not say it, but the recording sounds exactly like their voice. Adults are unsure whom to believe.

What might open the conversation. “We are going to take this seriously, but we also need to check whether this recording is what it appears to be. That is part of being fair to everyone.”

The Scam That Sounded Real

What adults might see. A parent receives a frantic phone call from what sounds exactly like their child's voice, saying they are in danger and need money immediately.

What may actually be happening. A scammer has cloned the child's voice from a public social media post. The call is AI-generated. The parent, hearing their child's voice in distress, acts on instinct.

What might open the conversation. "Let us set up a family code word — something only we know — so that if anyone ever calls sounding like one of us, we have a way to verify."

The Rumor Machine

What adults might see. A student is suddenly withdrawn, avoiding friends, and reluctant to come to school.

What may actually be happening. A fake image or video has been created showing the student in a compromising situation. Even though the content is fabricated, the student fears that people believe it is real. They feel powerless to prove otherwise and are terrified it will follow them forever.

What might open the conversation. "I have noticed you seem to be going through something. Whatever it is, I am on your side. You do not have to handle it alone."

The Identity Experiment

What adults might see. A teenager is spending time making AI-generated images and seems engaged in a creative project.

What may actually be happening. The teenager is altering photos of themselves — changing their face, body, or appearance — in ways that reflect genuine distress about how they look. The AI-generated version becomes the standard they measure themselves against, and the gap between the image and the mirror grows painful.

What might open the conversation. "Do you ever feel different about yourself after editing photos? Some people find that happens."

Different Ages, Different Vulnerabilities

These patterns help calibrate attention, not define rigid categories. Every young person is different.

Ages 9–11

Typical encounters	Mostly through gaming platforms, YouTube, and content shared by older siblings or peers. Less likely to seek out synthetic media, more likely to encounter it without context or explanation.
Key risks	Children this age tend to trust what they see on screens. Their natural skepticism is shallow and can be overridden by a trusted peer or an authoritative-looking source. They often lack the language to describe what they encountered or why it troubled them.
What adults might notice	Repeating claims from videos without questioning them. Confusion about whether something they saw was real. Distress they cannot articulate.
What support looks like	Concept-level foundation work: what is real, what is made up, and why it matters. The most important message for this age is “tell a trusted adult when something seems off.”

Ages 12–14

Typical encounters	Active on social media. This is the highest-risk group for intimate image abuse in schools. Financial sextortion targeting boys is heavily documented in this age band.
Key risks	Peer status is paramount, and the reputational damage from fake imagery is most acute. Impulse control is still developing; creating a deepfake can feel like a dare without understanding the consequences. Many victimized teens never tell a parent or trusted adult.
What adults might notice	Sudden mood changes after phone use. Withdrawal from friend groups. Reluctance to attend school. Anxiety about specific classmates.

What support looks like	Direct conversations about what deepfakes are and what happens to people who create or share them. Clear signals that disclosure will be met with support, not punishment or device removal.

Ages 15–18

Typical encounters	Full engagement across all platforms. Most likely to understand that deepfakes exist, but often underestimate how realistic current tools are. Celebrity deepfake scams target this group's growing financial independence.
Key risks	In surveys, many teenagers say a deepfake nude would feel worse than a real image shared — they grasp the permanence. In most places, creating AI-generated sexual images of classmates is now a criminal offense regardless of the creator's age.
What adults might notice	Overconfidence about ability to spot fakes. Casual attitude toward face-swap apps. Sharing synthetic media without considering the subject's consent.
What support looks like	Honest conversations about legal consequences. Media literacy skills including verification habits. Understanding that being technically capable of making something does not make it acceptable.

A note on neurodivergent young people. Young people with autism, ADHD, and other neurodevelopmental differences can face heightened risks around synthetic media. Literal interpretation of visual content, difficulty detecting social cues that signal deception, impulsive clicking or sharing, and challenges with abstract risk assessment can all increase vulnerability. Safety guidance works best when it is explicit, concrete, and rule-based rather than relying on instinct or social inference. Concrete examples and step-by-step response plans are more helpful than general advice to "be careful." Families and schools supporting neurodivergent young people may want to consider more active, transparent monitoring and stronger default privacy settings as a starting point for conversation.

Signals Worth Paying Attention To

These are not proof of a problem. They are invitations to be curious.

"I saw this video of [person] and I cannot believe they said that."

A young person reacting to media with shock or outrage may be encountering a deepfake — or may be about to share one.

“Everyone has seen it.”

When something is circulating fast and generating strong reactions, the speed itself is a signal. Viral content often outpaces verification.

Sudden withdrawal from school or social activities.

A student who was engaged and then suddenly pulls back may be dealing with the fallout of a deepfake — as a target, a witness, or someone who participated and is now afraid.

New secrecy around phone or computer use.

A shift toward hiding screens, deleting messages, or becoming defensive about online activity can mean many things. One of them is involvement — as creator, distributor, or target — in synthetic media.

Unexplained distress about appearance or reputation.

A young person who suddenly seems anxious about how they look or what people think of them may have encountered an altered image of themselves.

A child asking whether a recording or image is real.

Take this seriously. If they are asking, something has prompted the question.

How to Talk About Deepfakes With Young People

- Explain it simply: a deepfake is a fake image, video, or voice recording that can look or sound real. Seeing is no longer believing, so pausing is part of staying safe.
- Use one scenario as a “What would you do if…?” conversation. For example: What would you do if a forwarded image of a classmate showed up in a group chat?
- Practice verification questions together: Who posted this first? What do other sources say? Is there a reason someone would want us to believe this?
- Make the safety message clear before there is a crisis: “If someone ever makes a fake image or recording of you, it is not your fault and you can tell me.”
- Rehearse the first steps: pause, do not forward it, do not joke about it, and tell a trusted adult.
- Keep the tone steady. The goal is not to make young people afraid of every image or recording, but to help them slow down when something feels shocking, harmful, or too good to be true.

What Adults Can Do Right Now

At Home

- Talk about deepfakes before there is a crisis. Use a news story, a celebrity example, or a conversation about social media to introduce the idea that images and audio can be fabricated. Keep it matter-of-fact, not alarmist.
- Establish a family code word. Agree on a word or phrase that only family members know. If anyone ever calls sounding like a family member and asking for money or help, the code word is how you verify. Include grandparents and extended family.
- Reduce the raw material. Every public photo, video, and voice recording is potential source material for a deepfake. Help young people understand why privacy settings matter — not as a rule, but as protection.
- Make disclosure safe. Say it out loud: "If something like this ever happens to you or someone you know, you can tell me. I will not take your phone, and I will not blame you."

At School

- Name the issue. If your school has not talked about deepfakes with students, staff, and families, start now. Silence communicates that adults are unaware or unconcerned.
- Update policies. Acceptable-use policies should explicitly address AI-generated content, including synthetic intimate imagery. Make consequences clear and ensure they apply to creators and distributors, not to victims.
- Train staff. Teachers, counselors, and administrators need to know what deepfakes look like, how students encounter them, and how to respond when a student discloses. The first adult response matters enormously.
- Teach verification, not detection. Instead of investing in unreliable detection tools, teach students to ask: Where did this come from? Who posted it first? What do other sources say? Who benefits if this is believed?

If Something Goes Wrong

If a Young Person Is Targeted

- Believe them first. The most important thing a young person needs to hear is that you believe them and that this is not their fault. The content may look real. That does not change the fact that it was made without their consent.
- Preserve evidence carefully. Screenshot the URL, the username of the poster, and any surrounding messages — but do not save, print, or store the explicit content itself. Write a description of what the content shows and where you found it.
- Report to the platform. Every major platform has a mechanism for reporting non-consensual intimate imagery. Many now have specific pathways for AI-generated content.
- Report to the school and, where appropriate, to law enforcement. In most places, creating and distributing fake intimate imagery of a minor is a crime. Schools receiving federal funding have obligations under Title IX to address this as sexual harassment.
- Contact support organizations. The National Center for Missing and Exploited Children can be reached at 1-800-843-5678. For financial sextortion, contact the FBI. Crisis support is available through the Crisis Text Line (text HOME to 741741) and the 988 Suicide and Crisis Lifeline.
- Monitor wellbeing. The harm from a deepfake does not end when the content is removed. Watch for changes in mood, sleep, appetite, school attendance, and social withdrawal. Professional support from a counselor familiar with online harm can make a significant difference.

If a Young Person Created or Shared a Deepfake

- Take it seriously without catastrophizing. A young person who made a deepfake as a joke needs to understand the harm they caused and the legal risks they face. They also need adults who can help them make it right.
- Focus on harm, not just rules. “Do you understand what this feels like for the person in that image?” is a more powerful question than “Do you know you could be expelled?”
- Seek restorative approaches where possible. Accountability matters. So does the young person’s ability to learn from the experience and make a genuine repair.

What Schools Owe Their Communities

Schools cannot prevent every deepfake. But they can control how prepared they are when one surfaces. At a minimum, every school should have a clear, written plan that answers these questions:

- Who is the first point of contact when a deepfake is reported?
- What is the process for preserving evidence without storing explicit content?

- How will the school communicate with affected families?
- What support will be offered to the targeted student?
- What are the consequences for students who create or distribute synthetic intimate imagery?
- How will the school address the wider student body without amplifying the content?

Schools that have responded well to deepfake incidents share common traits: they acted quickly, believed the targeted student, communicated transparently with families, and treated the incident as a safeguarding matter rather than a technology glitch.

Small Steps Adults Can Take This Week

- Talk with one young person about how images, audio, and video can be faked, and why it helps to pause before reacting.
- Set up or review a family or household code word for verifying urgent voice calls.
- Agree at home, at school, or in a youth group not to forward shocking or explicit media until an adult has checked it.
- Choose one scenario from this guide and ask, “What would you do if this happened in your group chat, classroom, or family?”
- Check whether your school, youth program, or church can answer the questions in “What Schools Owe Their Communities.”
- Make sure every adult who works closely with young people knows where to report deepfake incidents: the school contact, platform tools, and the National Center for Missing & Exploited Children.

What Young People Deserve to Know

Adults sometimes hesitate to talk to young people about deepfakes, worried that the information itself is harmful. The opposite is true. Young people who understand what synthetic media is and how it works are better equipped to protect themselves and support their peers.

Every young person deserves to know that images, audio, and video can be fabricated convincingly. Seeing or hearing something is no longer enough to know it is real.

They deserve to know that if someone creates a fake image of them, it is not their fault — regardless of what photos they have posted or shared.

They deserve to know that creating a fake intimate image of another person is abuse, and in most places it is a crime — even if the creator is also a minor.

They deserve to know that telling a trusted adult is the single most important step when something goes wrong — and that the adults in their life can handle it.

And they deserve to know that the ability to question what you see, ask where something came from, and refuse to share something harmful — those are not technical skills. They are character.

Looking Ahead

The technology behind deepfakes will continue to improve. Detection will continue to fall behind. New tools will emerge, and the ones that exist now will become faster, cheaper, and more convincing.

That is precisely why this guide focuses on foundations rather than features. Trust, evidence, harm, and power are not tied to any particular app, platform, or algorithm. They will remain relevant regardless of how the technology changes.

What will also remain relevant is the relationship between a young person and the adults they trust. No detection tool, content filter, or school policy will ever replace a child who feels safe enough to say, "Something happened and I need help."

Build that safety now, before it is needed.

> *"No detection tool, content filter, or school policy will ever replace a child who feels safe enough to say, 'Something happened and I need help.'"*

Short Resources Note

Specific tools and platforms change quickly. Rather than listing apps or detection software that may be outdated by the time you read this, we recommend building relationships with a few organizations that stay current:

- Media literacy organizations focused on youth can help schools develop curriculum and conversation frameworks.

- School safety and child protection organizations offer reporting guidance, incident response templates, and staff training resources.
- National reporting channels — including resources for missing and exploited children, financial fraud, and crisis support — should be known to every adult in a young person's life before they are needed.

The foundations in this guide — trust, evidence, harm, and power — will outlast any single product or platform. Start there.

What Every Adult Should Know About Deepfakes and Young People

What deepfakes are.

AI can now create convincing fake video, images, and audio from a single photograph or a few seconds of someone's voice. The tools are free, fast, and require no technical skill. Detection tools do not work reliably.

How they reach young people.

The most common harm in schools is intimate image abuse: ordinary photos of classmates run through apps that generate fake explicit images, then shared through group chats and social media. Other forms include voice-cloning scams, fake videos of teachers, and AI-generated bullying content.

Four foundations that outlast the technology.

Trust — Seeing is no longer believing. Build the habit of pausing before reacting.
Evidence — Ask where something came from. Check other sources before accepting or sharing.
Harm — A fake intimate image causes real damage. Creating or sharing one is abuse, and in most places a crime.
Power — This technology follows existing power imbalances. Ask who benefits and who is hurt.

What to do right now at home.

- Talk about deepfakes before there is a crisis.
- Establish a family code word for verifying voice calls.
- Help young people minimize their public digital footprint.
- Make it clear that disclosure will be met with support, not punishment.

What to do right now at school.

- Name the issue in conversations with students, staff, and families.
- Update acceptable-use policies to address AI-generated content.
- Train staff on recognition and response.
- Teach verification habits, not detection tools.

If a young person is targeted.

Believe them first. Preserve evidence (URLs, usernames, descriptions — not the content itself). Report to the platform, the school, and law enforcement where appropriate. Contact the National Center for Missing and Exploited Children (1-800-843-5678) for cases involving minors. Monitor wellbeing and connect with professional support.

If a young person created or shared a deepfake.

Take it seriously. Focus on the harm caused, not just the rule broken. Help them understand what the targeted person experienced. Pursue accountability and, where possible, restorative repair.

One sentence to remember.

The adults who matter most in a young person’s life are not the ones who can spot a deepfake. They are the ones the young person feels safe enough to tell.

Founder's Note

This guide grows out of work focused on how young people and families are meeting AI in the background of ordinary life. With synthetic media, the change is not only technical. It alters what people trust, what counts as proof, and how quickly harm can move through a school, a friend group, or a family. The aim here is simple: to give non-technical adults steady language and habits so they can respond without panic, support young people well, and act early when something is wrong.

This brief is offered as an additional lens, not a replacement for safeguarding policies, legal advice, or the judgment of people who know their communities. The hope is to make adults more prepared, more credible, and easier to come to when something serious happens.

— Randy Kart
Founder, AstraEthica
astraethica.ai