Emerging Technologies & Career Advice

Peter Carragher, adapted from Trust & Safety Teaching Consortium

2026-04-18

Emerging Technologies & Career Advice

Peter Carragher, adapted from Trust & Safety Teaching Consortium

Learning Objectives

- Learn about AI/ML, AR/VR, and decentralized systems
- Examine how these technologies intersect with Trust & Safety
- Explore career pathways and opportunities

Learning Objectives

Today we will:

- Learn about AI/ML, AR/VR, and decentralized systems
- Examine how these technologies intersect with Trust & Safety
- Explore career pathways and opportunities

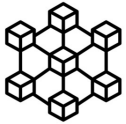
Emerging Technologies

Artificial Intelligence

AI is a rapidly growing technology that can be used to automate trust and safety processes.

Virtual Reality

AR/VR technologies are bringing people into augmented and immersive worlds.

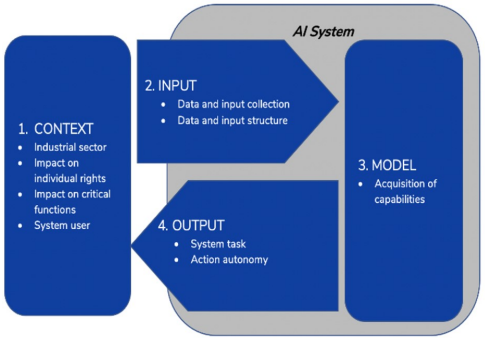


Web3 and the Fediverse

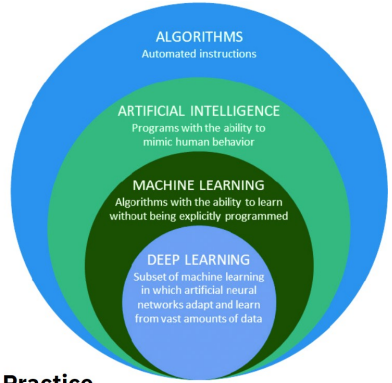
Web3 and blockchain technology decentralizes data, digital assets, and entire platforms.

Emerging Technologies
Artificial Intelligence
AI is a rapidly growing technology that can be used to automate trust and safety processes.
Virtual Reality
AR/VR technologies are bringing people into augmented and immersive worlds.
Web3 and the Fediverse
Web3 and blockchain technology decentralizes data, digital assets, and entire platforms.

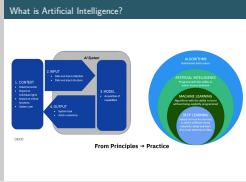
What is Artificial Intelligence?



OECD



From Principles → Practice



AI System/Definitions (left)

- OECD (image from: https://oecd.ai/en/wonk/cset-test-classification-ai-systems
- Other key definitions from
  - White House
  - NIST

AI Definitions (right)

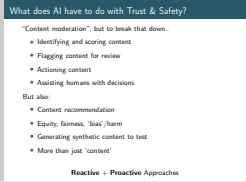
- CC license
- Image from this paper: https://www.researchgate.net/figure/sualization-of-algorithms-vs-artificial-intelligence-vs-machine-learning-vs-de fig1_339997962

What does AI have to do with Trust & Safety?

- Identifying and scoring content
- Flagging content for review
- Actioning content
- Assisting humans with decisions

- Content *recommendation*
- Equity, fairness, 'bias'/harm
- Generating synthetic content to test
- More than just 'content'

└What does AI have to do with Trust & Safety?



Example 1

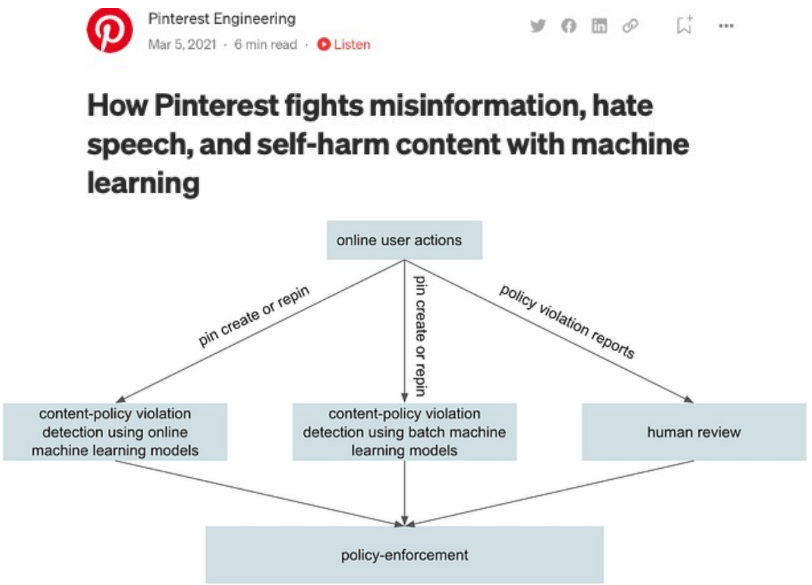
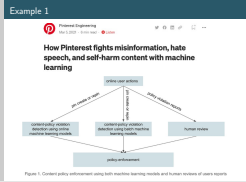


Figure 1. Content policy enforcement using both machine learning models and human reviews of users reports



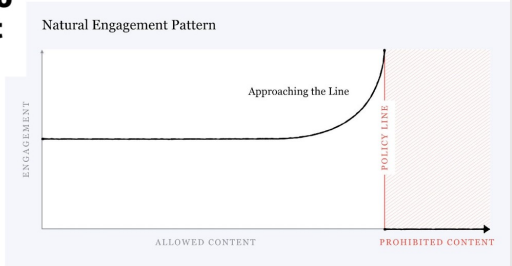
From:https://medium.com/pinterest-engineering/how-pinterest-fights-misinformation-hate-speech-and-self-harm-content-with-ma

Key points from post:

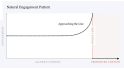
- Combination of ML + user reports
- Detect unsafe content before it's reported
- Combination of methods, data sources signals
- Signals conveyed from keywords, images, etc
- Key questions around measurement

Example 2

Facebook will change algorithm to demote "borderline content" that almost violates policies

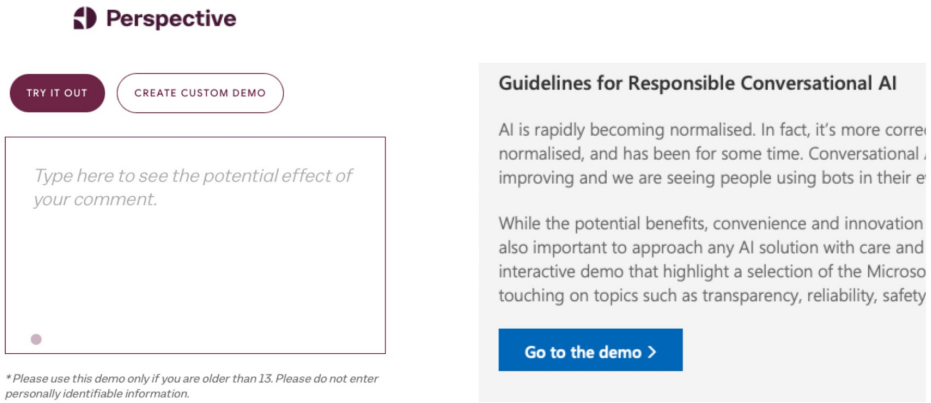


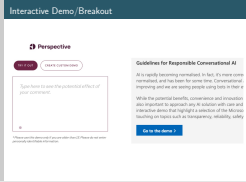
Example 2
Facebook will change algorithm to demote "borderline content" that almost violates policies



Usage of AI/ML to identify borderline content + prohibited content

Interactive Demo/Breakout





Try out one of these services/learning demos to learn more about:

- Toxicity/harmful speech detection - https://perspectiveapi.com/
- MSFT conversational responsible AI demos
- https://teachablemachine.withgoogle.com/

Discussion Questions

- What types of T&S-related topics can (and should) AI/ML based systems address?
- What is *not* well suited to be addressed by AI or automated methods in T&S?
- Policy development and enforcement often require judgements about context. What are important contextual signals that should be considered in human/machine environments?
- How would you measure the effectiveness of AI systems for T&S objectives?
- How should T&S practitioners think about differential impact across communities, including markets? Like those in the Global South, where training data may be less reflective of specific populations

Discussion Questions
- What types of T&S-related topics can (and should) AI/ML based systems address?
- What is *not* well suited to be addressed by AI or automated methods in T&S?
- Policy development and enforcement often require judgements about context. What are important contextual signals that should be considered in human/machine environments?
- How would you measure the effectiveness of AI systems for T&S objectives?
- How should T&S practitioners think about differential impact across communities, including markets? Like those in the Global South, where training data may be less reflective of specific populations

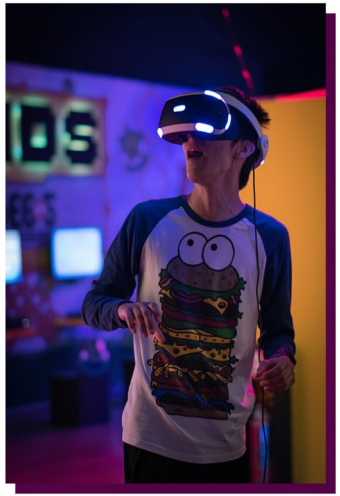
https://newpublic.substack.com/p/a-social-network-taxonomy

Defining Augmented and Virtual Reality

- **Augmented Reality**
  Overlaying digital information onto the real world, typically through a device like a smartphone or a headset
- **Virtual Reality**
  Fully immersive experiences in a simulated environment, typically through a dedicated VR headset

Defining Augmented and Virtual Reality
• Augmented Reality
Overlaying digital information onto the real world, typically through a device like a smartphone or a headset
• Virtual Reality
Fully immersive experiences in a simulated environment, typically through a dedicated

Example Harm Areas of AR and VR



- Harassment and bullying
- Inappropriate content
- Privacy concerns
- Age restrictions and child safety
- Accessibility and inclusivity

Example Harm Areas of AR and VR

In immersive and augmented spaces, long-standing areas of harm take on new dimensions:

- Harassment and bullying
- Inappropriate content
- Privacy concerns
- Age restrictions and child safety
- Accessibility and inclusivity

Discussion Questions

- How do trust and safety concerns differ between virtual reality (VR), augmented reality (AR), and traditional digital environments? Discuss the unique challenges and opportunities that immersive technologies present for content moderation.
- What are the ethical considerations surrounding the collection and use of user data in VR and AR environments? How can platforms ensure that they respect user privacy while promoting safety?
- Discuss the role of AI and machine learning in moderating VR and AR experiences. What are the benefits and limitations of these technologies, and how can they aid human moderation?
- How can VR and AR platforms foster a culture of empathy, respect, and inclusivity among users? Discuss the role of virtual etiquette, social norms, and digital citizenship in creating a safe and welcoming environment for all users.

Discussion Questions
- How do trust and safety concerns differ between virtual reality (VR), augmented reality (AR), and traditional digital environments? Discuss the unique challenges and opportunities that immersive technologies present for content moderation.
- What are the ethical considerations surrounding the collection and use of user data in VR and AR environments? How can platforms ensure that they respect user privacy while promoting safety?
- Discuss the role of AI and machine learning in moderating VR and AR experiences. What are the benefits and limitations of these technologies, and how can they aid human moderation?
- How can VR and AR platforms foster a culture of empathy, respect, and inclusivity among users? Discuss the role of virtual etiquette, social norms, and digital citizenship in creating a safe and welcoming environment for all users.

2026-04-18

Decentralized
Technologies

Blockchain, Cryptocurrencies, NFTs

2026-04-18

└Blockchain, Cryptocurrencies, NFTs

Blockchain, Cryptocurrencies, NFTs

Web3 and blockchain technology provides a decentralized way to structure platforms, data, and digital currencies.

It can also be used to create and manage ownership of digital assets that are used on tech platforms and even in-game experiences.

These novel technologies also introduce new abuse vectors that you may face as a Trust & Safety professional.

Blockchain, Cryptocurrencies, NFTs

- **Blockchain Technology** provides a decentralized and secure way to store data and conduct transactions.
- **Key Attributes**
  - Decentralization
  - Encryption
  - Consensus mechanisms
- **Cryptocurrency** Digital or virtual currency that uses cryptography for security and operates on a decentralized network.
- **NFTs** 'Non-Fungible Tokens', or a digital asset with a unique identifier.

Blockchain, Cryptocurrencies, NFTs
- **Blockchain Technology** provides a decentralized and secure way to store data and conduct transactions.
- **Key Attributes**
  - Decentralization
  - Encryption
  - Consensus mechanisms
- **Cryptocurrency** Digital or virtual currency that uses cryptography for security and operates on a decentralized network.
- **NFTs** 'Non-Fungible Tokens', or a digital asset with a unique identifier.

Blockchain, Cryptocurrencies, NFTs

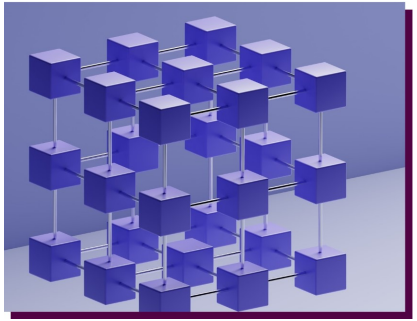
- **Cryptocurrency:** Digital or virtual currency that uses cryptography for security and operates on a decentralized network.
- **NFTs:** 'Non-Fungible Tokens', or a digital asset with a unique identifier.

Blockchain, Cryptocurrencies, NFTs

Example Technologies:

- **Cryptocurrency:** Digital or virtual currency that uses cryptography for security and operates on a decentralized network.
- **NFTs:** 'Non-Fungible Tokens', or a digital asset with a unique identifier.

Blockchain, Cryptocurrencies, NFTs

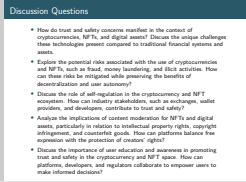


- Decentralization can raise major harm vectors, including:
  - Account Stealing and other platform integrity vulnerabilities
  - Financial Fraud and Scams
  - Inauthentic Behavior

Blockchain, Cryptocurrencies, NFTs
Decentralization can raise major harm vectors, including:
Account Stealing and other platform integrity vulnerabilities
Financial Fraud and Scams
Inauthentic Behavior

Discussion Questions

- How do trust and safety concerns manifest in the context of cryptocurrencies, NFTs, and digital assets? Discuss the unique challenges these technologies present compared to traditional financial systems and assets.
- Explore the potential risks associated with the use of cryptocurrencies and NFTs, such as fraud, money laundering, and illicit activities. How can these risks be mitigated while preserving the benefits of decentralization and user autonomy?
- Discuss the role of self-regulation in the cryptocurrency and NFT ecosystem. How can industry stakeholders, such as exchanges, wallet providers, and developers, contribute to trust and safety?
- Analyze the implications of content moderation for NFTs and digital assets, particularly in relation to intellectual property rights, copyright infringement, and counterfeit goods. How can platforms balance free expression with the protection of creators' rights?
- Discuss the importance of user education and awareness in promoting trust and safety in the cryptocurrency and NFT space. How can platforms, developers, and regulators collaborate to empower users to make informed decisions?



The Fediverse

The Fediverse

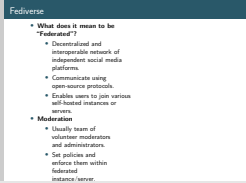
The advent of new technologies includes converging with what has come so far. Platforms still create opportunities to create content, but now they're federated across different servers.

Content, games, art, and currency is built on the blockchain. Ownership is in the hands of users, granting greater user freedoms and ownership, and challenging traditional models of keeping platforms safe.

In this section, we go over some of the promises and challenges of decentralized technologies.

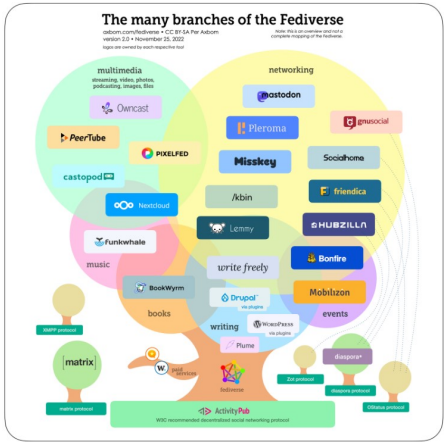
Fediverse

- **What does it mean to be "Federated"?**
  - Decentralized and interoperable network of independent social media platforms.
  - Communicate using open-source protocols.
  - Enables users to join various self-hosted instances or servers.
- **Moderation**
  - Usually team of volunteer moderators and administrators.
  - Set policies and enforce them within federated instance/server.



Fediverse and Content Moderation

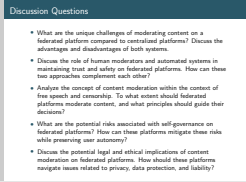
- Fediverse communities are governed by administrators or moderators
- Similar to internet forums, community platforms (e.g. Facebook Pages, Reddit subreddits, Discord servers)
- What's different? There's a limited centralized Trust & Safety function



Fediverse and Content Moderation
- Fediverse communities are governed by administrators or moderators
- Similar to internet forums, community platforms (e.g. Facebook Pages, Reddit subreddits, Discord servers)
- What's different? There's a limited centralized Trust & Safety function

Discussion Questions

- What are the unique challenges of moderating content on a federated platform compared to centralized platforms? Discuss the advantages and disadvantages of both systems.
- Discuss the role of human moderators and automated systems in maintaining trust and safety on federated platforms. How can these two approaches complement each other?
- Analyze the concept of content moderation within the context of free speech and censorship. To what extent should federated platforms moderate content, and what principles should guide their decisions?
- What are the potential risks associated with self-governance on federated platforms? How can these platforms mitigate these risks while preserving user autonomy?
- Discuss the potential legal and ethical implications of content moderation on federated platforms. How should these platforms navigate issues related to privacy, data protection, and liability?



2026-04-18

Career Advice

Careers in Trust & Safety

2026-04-18

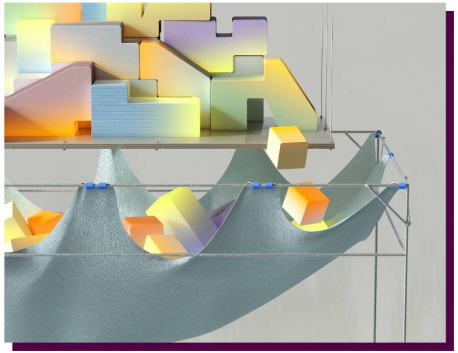
└Careers in Trust & Safety

Careers in Trust & Safety

In industry, Trust & Safety professionals are responsible for ensuring the safety of users and clients in the policies they form, products they build, and stories they tell in telling an organization's story to the world.

As new technologies emerge and novel forms of harm proliferate, the impact of Trust & Safety work will be more important than ever.

Trust & Safety Roles in Tech

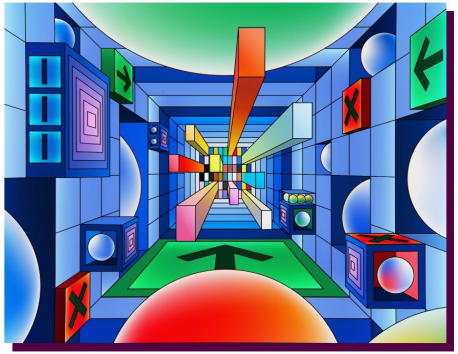


- **Legal:** Guide company across teams on risk mitigation, develop responses to regulatory bodies and law enforcement.
- **Policy:** Design content policies, performing risk assessments for product teams, and forming partnerships with civil society and governments.
- **Operations:** Enforce policy through scaled operational programs. Dedicated teams across subject-matter expertise, program and project management.

Trust & Safety Roles in Tech

- **Legal**: Guide company across teams on risk mitigation, develop responses to regulatory bodies and law enforcement.
- **Policy**: Design content policies, performing risk assessments for product teams, and forming partnerships with civil society and governments.
- **Operations**: Enforce policy through scaled operational programs. Dedicated teams across subject-matter expertise, program and project management.

Trust & Safety Roles in Tech



- **Product & Engineering:** Builds internal and consumer-facing products, including T&S tooling to safety features
- **Data Science:** Analyzes policy violations and their impact on platforms, develops measurement and detection methods, and create strategies to address harms.
- **Marketing & Communications:** Develops narrative strategies for showcasing organization's safety commitments. Drives content strategies and user research efforts.

Trust & Safety Roles in Tech

- **Product & Engineering:** Builds internal and consumer-facing products, including T&S tooling to safety features
- **Data Science:** Analyzes policy violations and their impact on platforms, develops measurement and detection methods, and create strategies to address harms.
- **Marketing & Communications:** Develops narrative strategies for showcasing organization's safety commitments. Drives content strategies and user research efforts.

Trust & Safety Partnerships

2026-04-18

└Trust & Safety Partnerships

Working in Trust & Safety doesn't have to be limited to working at a tech company.

In fact, tech companies need extensive collaborations and partnerships across industry, academic, civil society, and government partners.

Industry Partnerships

- Fellow tech companies and industry working groups
- Coordinate on operations and investigations
- Collaborate in industry working groups (e.g. TSPA) and share best practices
- Collaborate on guiding regulators

Industry Partnerships

- Fellow tech companies and industry working groups
- Coordinate on operations and investigations
- Collaborate in industry working groups (e.g. TSPA) and share best practices
- Collaborate on guiding regulators

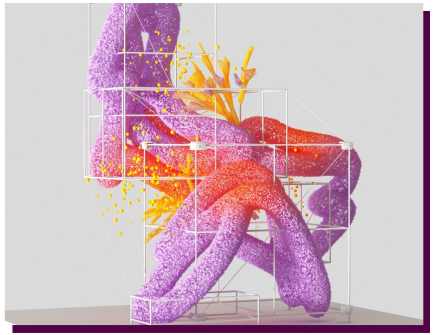
Academic Partnerships

- Partner on user research and social impact
- Guide product teams
- Audit content policies and public policy initiatives
- Example partners: academic labs, PhD and Masters students, specialized researchers and research projects

Academic Partnerships

- Partner on user research and social impact
- Guide product teams
- Audit content policies and public policy initiatives
- Example partners: academic labs, PhD and Masters students, specialized researchers and research projects

Civil Society Partnerships



- Create guiding frameworks for policy, product, data privacy, Trust & Safety operations
- Co-create education resources and industry reports to guide industry awareness of issues
- Example Partners: subject-matter focused (e.g. child safety), product-based (e.g. organizations developing AI product and policy frameworks)

Civil Society Partnerships

- Create guiding frameworks for policy, product, data privacy, Trust & Safety operations
- Co-create education resources and industry reports to guide industry awareness of issues
- Example Partners: subject-matter focused (e.g. child safety), product-based (e.g. organizations developing AI product and policy frameworks)

Governments

- Regulate platform content moderation practices and safety products
- Issue law enforcement inquiries (e.g. subpoenas, exigent requests)
- Example partners: local, provincial, national, and international governing bodies, law enforcement agencies

Governments

- Regulate platform content moderation practices and safety products
- Issue law enforcement inquiries (e.g. subpoenas, exigent requests)
- Example partners: local, provincial, national, and international governing bodies, law enforcement agencies

Readings / References

- Blockchain icon created by Good Ware
- Brain icon created by Freepik
- VR icon created by Nikita Golubev
- Fediverse map by Per Axbom, https://axbom.com/fediverse/
- Career Advice section art from DeepMind's Visualising AI project, https://visualisingai.deepmind.com

2026-04-18
Readings / References