Generative AI Media and Art

Examination of ethical, legal, and societal issues in generative AI for media and art, including biases in training data, copyright concerns in model training, regulatory proposals, impacts on creative professions, risks from deepfakes and synthetic media, and value embeddings in prompting systems.

Ghassem Tofighi

Contents

Overview

This lesson addresses the ethical, legal, and societal aspects of generative AI in media and art. Topics include biases in training data and outputs, copyright issues related to training processes, regulatory frameworks for large models, effects on creative industries, risks from deepfakes and synthetic media, and value embeddings in prompting systems. The lesson draws on required readings to provide analysis of these dimensions.

Learning Objectives

- Explain sources of bias in generative AI and their links to training data composition.
- Assess copyright challenges in training generative models on protected works.
- Describe proposed regulatory mechanisms for generative AI systems.
- Analyze labor market changes in creative fields due to generative AI.
- Identify harms from deepfakes in areas such as harassment and misinformation.
- Examine how prompting influences model outputs and embeds cultural values.

Motivation

Generative AI produces content that intersects with human creativity, raising questions about fairness, ownership, and societal impact. Understanding these issues informs the development and application of such technologies.

Ethical Implications of Generative AI

Generative AI systems introduce distinct ethical concerns that arise at multiple stages: data collection, model training, deployment, and use.

1. Consent and Privacy in Data Collection

Most large-scale generative models are trained on massive web-scraped datasets containing personal images, text, voices, and artworks.
Individuals whose content appears in training data typically did not provide explicit consent for use in commercial AI systems.
Right of publicity and data protection laws (e.g., GDPR Article 6, CCPA) raise questions about the lawful basis for processing.

Example: Facial images of private individuals were used to train face-generation models without permission.

2. Perpetuation and Amplification of Harmful Stereotypes

Training data reflects historical and current societal biases.
Models learn and reproduce associations that disadvantage marginalized groups.

Example: Text-to-image models generate doctors predominantly as male and nurses as female when no gender is specified.

Example: Language models associate certain ethnic groups with negative stereotypes more frequently than baseline rates in curated corpora.

3. Economic Concentration and Power Asymmetries

The infrastructure required to train frontier generative models is accessible to only a small number of organizations.
Control over model weights, fine-tuning access, and API pricing concentrates economic and cultural influence.
Independent artists, small studios, and local media organizations face barriers to participation.

Example: A handful of companies determine safety alignments and content filters applied globally.

4. Erosion of Human Creative Labour Value

Generative tools enable the rapid production of content that previously required skilled human input.
Displacement risk is highest in routine, mid-tier creative work (stock photography, commercial copywriting, entry-level illustration).
When human-created works are used to train replacement systems, creators lose both current income and future bargaining power.

Example: A freelance illustrator discovers AI-generated images in the exact style they developed over years being sold at a fraction of their rates.

5. Environmental and Resource Justice

Training and inference of large generative models consume substantial electricity and water.
Energy demand is comparable to that of small cities during training runs.
Carbon footprint disproportionately affects regions already impacted by climate change.

Example: A single training run of a 100B+ parameter multimodal model can emit hundreds of tons of CO_2 equivalent.

6. Non-Consensual Representations and Dignity Harms

Generative systems enable the creation of realistic but fabricated content involving real people.
Non-consensual intimate imagery (deepfake pornography) constitutes a form of gender-based violence.
Fabricated depictions can damage reputation, particularly when targeting public figures, journalists, or activists.

Example: Explicit deepfakes of actresses and politicians circulated without consent, often used for harassment or extortion.

7. Epistemic and Democratic Risks

Widespread synthetic media undermines shared epistemic foundations.
The ability to produce convincing false evidence at scale weakens trust in visual and audio records.
Coordinated disinformation campaigns become cheaper and more plausible.

Example: Synthetic video of a political leader making inflammatory statements spreads hours before fact-checks can respond.

8. Value Alignment and Cultural Domination

Reinforcement learning from human feedback (RLHF) and content filters embed the preferences of annotators and safety teams.
These teams are often concentrated in high-income countries and specific demographic groups.
Resulting alignments may reflect Western, English-centric, corporate-friendly norms.

Example: Models refuse to generate content about certain political topics while permitting others, reflecting annotator consensus rather than universal values.

Summary of Ethical Dimensions

Dimension	Core Ethical Question	Primary Stakeholders Affected	Example Mitigation Approaches
Consent & Privacy	Was permission obtained for training use?	Individuals in datasets	Opt-out mechanisms, synthetic data
Bias & Stereotypes	Does the system reproduce societal harms?	Marginalized groups	Targeted debiasing, diverse annotation
Economic Concentration	Who controls frontier capabilities?	Independent creators, smaller firms	Open-weight models, public infrastructure
Labour Displacement	How is creative work valued in an AI-augmented world?	Working artists, writers, designers	Revenue sharing, licensing frameworks
Environmental Impact	Is the carbon cost justified?	Future generations, climate-vulnerable regions	Efficient architectures, carbon-aware training
Non-Consensual Content	Can real people be protected from harmful fakes?	Individuals targeted by deepfakes	Watermarking, detection mandates, prohibitions
Epistemic Trust	Can shared reality survive synthetic media?	Democratic institutions, journalism	Provenance standards, media literacy
Cultural Value Alignment	Whose norms are encoded as default?	Global cultural diversity	Multilingual/diverse RLHF, customizable filters

These ethical implications require analysis across technical, legal, economic, and cultural dimensions rather than purely technical solutions.

Biases in Generative AI and Training Data

Generative models learn statistical patterns directly from their training corpora. When these corpora contain imbalances, exclusions, or stereotypical associations, the learned distributions are reflected in model outputs.

Sources of Bias

1. **Historical and Cultural Imbalance in Visual Datasets**
   Large image-text datasets are dominated by content from Western, high-income countries.
   Art-historical collections used in training often under-represent non-Western traditions, female artists, and artists of colour.

   Example: Generated classical portraits show predominantly European facial features and attire even when prompts request “a classical portrait from any culture”.

2. **Gender and Role Stereotypes**
   Occupational and social roles in training data follow historical gender distributions.

Example: When no gender is specified, text-to-image models produce male-coded images for "CEO", "scientist", "engineer" and female-coded images for "nurse", "teacher", "secretary".

Example: Generated images of "a successful business person" default to men in suits far more frequently than women or non-binary individuals.

3. **Racial and Ethnic Skew**
   Beauty, professionalism, and positive attributes are disproportionately associated with lighter skin tones in training data.

   Example: Prompts for "beautiful person", "attractive face", or "professional headshot" yield lighter skin tones and Eurocentric features at rates higher than global population distribution.

4. **Linguistic and Textual Stereotypes**
   Language models trained on web text reproduce correlations present in large corpora, including racist, sexist, ableist, and homophobic associations.

   Example: Completing sentences such as "People from [country X] are..." frequently produces negative stereotypes that align with prevalent online discourse rather than factual distributions.

   Example: Associating certain accents, dialects, or languages with lower intelligence or lower social status when generating character descriptions.

Mechanisms of Bias Propagation

- **Representation bias**: Some groups appear less frequently or in narrower roles → lower probability of generating diverse outputs.
- **Co-occurrence bias**: Certain attributes (skin tone, gender, occupation) appear together more often → models learn spurious correlations.
- **Amplification effect**: Small imbalances in training data become exaggerated in generation because models sample from learned distributions.
- **Compression effect**: Rare identities or styles are collapsed toward majority modes during generation.

Observable Patterns in Outputs

Prompt Type	Common Bias Pattern	Typical Output Characteristics
Neutral profession	Gender skew toward historical norms	Male doctors, female nurses without prompt specification
Aesthetic / beauty	Racial and skin-tone bias toward Eurocentric standards	Lighter skin, specific facial features over-represented
Cultural / historical figure	Western-centric default	European-style clothing and settings for generic prompts
Emotional / character description	Stereotypical personality-trait associations	Angry = certain ethnic groups; nurturing = female-coded
Artistic style transfer	Over-representation of canonical Western artists	Van Gogh / Picasso styles dominate over non-Western ones

Partial Mitigation Strategies and Their Limits

- **Data filtering**: Removal of explicitly harmful content reduces severity but does not address structural under-representation.
- **Re-weighting / re-sampling**: Increases frequency of minority examples but can introduce new artifacts or over-correction.
- **Adversarial debiasing**: Trains models to minimize certain associations but often trades one bias for another.
- **Post-processing filters**: Catches blatant stereotypes in outputs but misses subtle statistical skews.
- **Diverse annotation for alignment**: RLHF with broader annotator pools helps but remains limited by who can participate in annotation work.

Bias in generative systems is not an isolated technical problem; it reflects the composition of training data, which in turn mirrors historical power structures, collection practices, and cultural visibility.

Generative AI Governance and Regulation

Governance and regulation of generative AI address risks that emerge at scale while attempting to preserve technical innovation and access. Approaches range from self-regulation to binding legal obligations.

Core Regulatory Challenges

- Models with billions of parameters can generate content across text, image, audio, and video domains.
- Harms occur downstream from deployment and are difficult to predict during training.
- Rapid capability improvements outpace traditional regulatory cycles.
- Global deployment creates jurisdictional fragmentation.

Proposed Regulatory Layers

1. **Product Safety and Risk Classification**
   Treat high-capability generative models as regulated products similar to medical devices or aviation software.
   Assign risk tiers based on parameter count, training compute, or demonstrated misuse potential.

   Example: Models above 10^{25} FLOPs of training compute classified as systemic-risk models requiring mandatory third-party audits.

2. **Transparency and Documentation Obligations**
   Require disclosure of key information to enable external scrutiny.

   - Model architecture and parameter count
   - Training data sources and filtering methods
   - Energy consumption and carbon footprint
   - Safety evaluation results
   - Change logs for fine-tuning and alignment

   Example: Mandatory model cards or datasheets published before public release, including known failure modes and bias measurements.

3. **Pre-Deployment Risk Assessment and Red-Teaming**
   Require systematic evaluation of misuse vectors before models are made publicly accessible.

- Adversarial testing for jailbreaks
- Assessment of synthetic media generation quality
- Evaluation of harmful content generation rates
- Stress-testing under high-risk prompts

Example: Independent red-team reports must be submitted to regulators demonstrating that non-consensual intimate imagery generation has been suppressed below a defined threshold.

4. **Liability and Accountability Mechanisms**
   Establish clear responsibility chains for harms caused by model outputs.

   - Developer liability for foreseeable misuse
   - Platform liability for hosted content generated by third-party models
   - User liability in cases of intentional malicious use

   Example: Strict liability for developers when models produce non-consensual deepfake pornography even after safety mitigations.

5. **Use-Case Restrictions and Prohibitions**
   Ban or heavily restrict applications with high, non-compensable harm.

   - Non-consensual intimate deepfakes
   - Real-time impersonation in electoral contexts
   - Child sexual abuse material generation
   - Biometric identification spoofing tools

   Example: Criminal penalties for creating or distributing AI-generated non-consensual explicit content depicting identifiable individuals.

Comparison of Major Governance Approaches

Approach	Key Mechanism	Responsible Entity	Enforcement Strength	Innovation Impact
Self-regulation / Voluntary commitments	Industry codes of conduct, safety pledges	Companies	Low	Minimal
Transparency-focused	Mandatory reporting and model cards	Regulators / independent bodies	Medium	Moderate
Risk-based product safety	Tiered obligations based on capability	Government agencies	High	Significant
Liability-driven	Legal responsibility for downstream harms	Courts / regulators	High	High
Use-case prohibitions	Categorical bans on high-risk applications	Legislatures	Very high	Targeted

Emerging International and Regional Frameworks

- **European Union AI Act** classifies general-purpose AI models with systemic risk thresholds and imposes transparency, risk assessment, and evaluation duties.
- **United States** combines executive orders (requiring safety testing for frontier models) with proposed legislation targeting deepfakes and watermarking.
- **China** enforces content alignment with state values through pre-approval of generative services.
- **Voluntary industry commitments** include watermarking synthetic content, publishing safety reports, and restricting certain misuse vectors.

Implementation Tensions

- **Speed vs. safety**: Pre-release evaluations delay access to new capabilities.
- **Open vs. closed models**: Open-weight models complicate control after release.
- **Global vs. local rules**: Fragmented regulation creates compliance complexity.
- **Measurement challenges**: Defining thresholds for acceptable risk or bias remains contested.

Effective governance requires balancing ex-ante controls (before release) with ex-post accountability (after harms occur), while recognizing that no single mechanism fully addresses the adaptive nature of generative AI misuse.

Training Data, Copyright, and Creative Labour

The use of large-scale datasets to train generative models raises fundamental legal and economic questions about ownership, fair compensation, and the future value of human creative work.

Core Legal Issues in Training Data Usage

1. **Reproduction Right**
   Copying protected works into a training dataset constitutes reproduction under most copyright regimes.

   Example: Downloading and storing millions of images from an artist's website or social media feed to train a style-transfer model reproduces those works in digital form.

2. **Preparation of Derivative Works**
   Models learn compressed representations of input works. Generated outputs that closely resemble specific training examples may be considered unauthorized derivatives.

   Example: An AI system trained on a specific illustrator's portfolio consistently produces new images that retain distinctive line quality, colour palette, and compositional habits of that illustrator.

3. **Fair Use / Fair Dealing Analysis**
   In jurisdictions that recognize fair use (primarily the United States), four factors are weighed:

   - Purpose and character of the use (commercial vs. transformative)
   - Nature of the copyrighted work (creative vs. factual)
   - Amount and substantiality of the portion used
   - Effect of the use upon the potential market or value of the original

   Commercial training of generative models often scores poorly on the first and fourth factors.

Example: A company trains a model on a database of professional photographs and then licenses the model for commercial image generation, directly competing with the market for stock photography and commissioned work.

Economic and Labour Dimensions

- **Input Without Compensation**
  Creators whose works are scraped and used receive no payment or credit for contributing to the training process.
- **Output Market Substitution**
  Generated content competes directly with human-created work in commercial contexts.

  Example: AI-generated book cover illustrations, advertising visuals, and social media graphics reduce demand for entry-level and mid-tier freelance designers.
- **Style Imitation and Loss of Scarcity**
  When a model can reliably imitate an artist's recognizable style, the economic value of developing and maintaining that style decreases.

  Example: A distinctive painterly style that took decades to develop becomes replicable at near-zero marginal cost after the model is trained.
- **Precarisation of Creative Labour**
  Shift from commission-based or royalty-based income toward gig-style prompt engineering or data annotation work.

  Example: Former concept artists now annotate training data or test model outputs for safety and quality, often under short-term contracts with limited bargaining power.

Practical Scenarios and Implications

Scenario	Copyright Question Raised	Labour Market Effect	Typical Stakeholder Positions
Scraping public Instagram posts for training	Does posting online imply license to train AI?	Reduced commissions for similar visual styles	Artists: no; Platforms: yes
Training on scanned books / digitized art	Is digitization + ML training fair use?	Displacement in illustration and design markets	Publishers/artists: no; AI companies: yes
Model outputs sold as stock images	Are outputs infringing derivatives?	Direct price competition in stock libraries	Stock agencies: concerned; AI providers: no issue
Fine-tuning on a single artist's portfolio	Clearer case of unauthorized derivative training	Severe impact on that artist's market value	Artist: strong claim; Fine-tuning user: weaker
Open datasets built from licensed material	Reduced legal risk if licensing is documented	Still competes with original creators' markets	More acceptable to some creators if compensated

Emerging Responses and Proposals

- **Opt-out registries** allowing creators to block their work from future training runs
- **Licensing marketplaces** for training data with revenue sharing
- **Compulsory licensing schemes** similar to music performance rights
- **Training-data provenance requirements** under proposed regulations
- **Collective bargaining** by creator guilds for remuneration from AI companies

The tension between enabling rapid innovation through broad data access and preserving economic incentives for human creative production remains unresolved in most jurisdictions.

Deepfakes and Synthetic Media: Harassment, Misinformation, and Manipulation

Synthetic media technologies enable the automated creation of highly realistic but fabricated audio, images, and video. When applied to real people or events, these capabilities introduce serious risks across personal, social, and institutional domains.

Core Technical Capabilities

- Face-swapping and lip-sync video synthesis
- Voice cloning from short audio samples
- Full-body motion transfer and reenactment
- Text-to-video generation of realistic scenes
- Audio-visual desynchronization for fabricated dialogue

These methods have improved rapidly, with generation quality now approaching photorealism in controlled settings.

Primary Harm Vectors

1. **Non-Consensual Intimate Imagery**
   Generation of explicit content depicting real individuals without their consent.

   Example: Deepfake pornography created by swapping the face of a known person onto an existing adult video, then distributed on pornographic websites or sent directly to the target or their contacts.

   Example: High-school or university students targeted with fabricated nude or sexual images shared within peer networks.

2. **Harassment and Reputation Damage**
   Fabricated media used to humiliate, intimidate, or discredit individuals.

   Example: Synthetic video of a public figure appearing intoxicated or making offensive statements circulated during a political campaign or workplace dispute.

   Example: Deepfake audio of a journalist confessing to fabricating stories, sent to their editors and posted online.

3. **Political Misinformation and Election Interference**
   Synthetic media used to simulate statements or actions by political actors.

   Example: A fabricated video of a candidate appearing to make inflammatory remarks about a minority group, released days before an election.

Example: Cloned voice of a sitting official issuing false orders or contradictory policy statements.

4. **Fraud and Impersonation**
   Real-time or near-real-time synthetic media used to deceive individuals or organizations.

   Example: Voice cloning used in CEO fraud scams, where the cloned voice of an executive instructs a finance employee to transfer large sums to a new account.

   Example: Video call impersonation during remote job interviews or identity verification processes.

5. **Erosion of Evidentiary Trust**
   Widespread availability of convincing fakes undermines confidence in visual and audio evidence.

   Example: Genuine footage of police violence dismissed as "probably a deepfake" by officials or online audiences.

   Example: Courtroom video evidence challenged on the basis that it could have been synthetically generated or altered.

Detection and Countermeasures Landscape

Countermeasure Type	Description	Current Effectiveness	Main Limitations
Forensic signal analysis	Pixel-level artifacts, compression inconsistencies, physiological signals	Moderate	Rapidly obsoleted by new generation methods
Watermarking / provenance	Embedded imperceptible markers during creation	Low in practice	Easily stripped or not adopted at source
Media authentication APIs	C2PA / Content Credentials standards	Emerging	Adoption remains low outside major platforms
Classifier-based detection	Trained models to distinguish real vs. synthetic content	Variable	High false positives/negatives; domain shift
Contextual verification	Cross-referencing metadata, source chain, consistency with known facts	High when applicable	Slow; does not scale to viral content
Legislative prohibitions	Criminalization of non-consensual intimate deepfakes	High deterrent value	Enforcement difficult for anonymous creators

Timeline of Escalation Risk

- **2017–2019**: Early face-swap tools; mostly low-quality, limited to static images or short clips
- **2020–2022**: Voice cloning reaches usable quality from minutes of audio; video deepfakes improve realism
- **2023–2025**: Text-to-video models produce coherent multi-second clips; real-time voice and face synthesis becomes feasible

- **Current frontier**: Multimodal models generate synchronized audio-visual content from text prompts in seconds

Societal and Institutional Implications

- Shift from "seeing is believing" to systematic skepticism of media evidence
- Increased burden on journalists, fact-checkers, and courts to authenticate content
- Pressure on platforms to implement proactive detection and removal (with corresponding free-expression concerns)
- Growing demand for digital provenance infrastructure that survives post-production editing

The asymmetry between generation speed/cost and reliable detection/verification cost creates a structural advantage for malicious actors in the short to medium term.

Summary

This lesson examines ethical, legal, and societal issues of generative AI in media and art. It covers ethical concerns such as lack of consent in data use, stereotype amplification, economic concentration, labour displacement, environmental impact, non-consensual representations, epistemic risks, and cultural value alignment. Biases stem from historical imbalances, gender/role stereotypes, racial skews, and linguistic associations, propagated through representation, co-occurrence, amplification, and compression effects. Governance proposals include product safety tiers, transparency obligations, pre-deployment risk assessment, liability rules, and use-case prohibitions. Training data and copyright issues involve reproduction rights, derivative works, fair use limitations, uncompensated use, market substitution, and precarisation of creative labour. Creative professions face task substitution, deskilling, opportunity polarization, and new precarious roles. Deepfakes and synthetic media create risks of non-consensual intimate content, harassment, political misinformation, fraud, and eroded trust in evidence. Prompting and power dynamics show how reinforcement learning, safety filters, language biases, and access differences embed specific cultural norms and values in models.