AI and Copyright

Training AI Models on Copyrighted Data

Security Course

Today's Agenda

1. How AI Models Learn from Data
2. The Scope of the Copying
3. Current Lawsuits and Legal Challenges
4. Fair Use Analysis Applied to AI
5. The Memorization Problem
6. Competing Policy Arguments
7. International Approaches
8. Future Directions

Recap: What We've Learned

Lecture 13 - AI and Privacy:

- LLMs pose unique privacy risks
- Memorization of training data
- User disclosure behaviors

Lecture 14 - Copyright and Fair Use:

- Four-factor fair use test
- Google v. Oracle (transformative use)
- Copyright detection and over-blocking

Today: Apply copyright law to AI training

How AI Models Learn

Traditional programming:

```
If user says "hello" → respond "hi"
If user says "goodbye" → respond "bye"
```

Machine learning:

```
1. Collect millions of examples (training data)
2. Model finds statistical patterns
3. Generate new outputs based on patterns learned
```

Key insight: Modern AI models are trained on massive amounts of data

What Data Are AI Models Trained On?

Large Language Models (GPT, Claude, etc.):

- Books (millions of published works)
- News articles and journalism
- Web pages and blogs
- Code repositories (GitHub, etc.)
- Scientific papers
- Social media posts

Image Models (Stable Diffusion, Midjourney, DALL-E):

- Billions of images from the internet
- Stock photo databases

The Scale of Copying

Common Crawl:

- Web scraping dataset
- Billions of web pages
- Used to train GPT-3, GPT-4, and others
- Includes copyrighted news, blogs, books

Books3:

- Dataset of ~200,000 books
- Scraped from pirate libraries
- Used to train many LLMs
- Includes recent bestsellers, textbooks

Is This Different From Google v. Oracle?

Similarities:

- Both involve copying for new technological purpose
- Both claim transformative use
- Both serve different markets than original

Differences:

- Scale: Billions of works vs. 11,500 lines of code
- Reproduction: AI can regenerate training data
- Purpose: Statistical learning vs. compatibility
- Market impact: Potentially substitutes for originals

Question: Do these differences matter for fair use analysis?

Current Lawsuits: The New York Times v. OpenAI

Filed: December 2023

Plaintiffs: The New York Times

Defendants: OpenAI and Microsoft

Claims:

- OpenAI trained models on millions of NYT articles
- ChatGPT can reproduce substantial portions of articles
- Substitutes for NYT subscription
- Harms licensing market

Damages: Billions in statutory damages + actual damages

NYT's Evidence

Complaint showed examples of:

- ChatGPT reproducing first paragraphs of NYT articles verbatim
- Outputting paywalled content without attribution
- Summarizing recent articles (potential subscription substitute)

NYT's argument:

- This isn't transformative use
- It's commercial exploitation
- Market substitution for subscription revenue
- Licensing alternative exists

OpenAI's Response

Arguments:

- Training is transformative (learning patterns, not copying)
- Output is new and original, not reproduced
- Fair use supports research and innovation
- Similar to Google Books (snippets for search)
- Any reproduction is rare "hallucination," not intentional

Motion to dismiss some claims:

- Argued training data can't be directly accessed
- Model weights don't contain "copies"
- Outputs are generated, not retrieved

The Authors Guild Lawsuits

Multiple lawsuits filed by groups of authors:

1. **Authors Guild v. OpenAI (Sept 2023)**
    - Authors: John Grisham, George R.R. Martin, Jodi Picoult, 17 others
    - Claims: Books used in training without permission
    - Books3 dataset identified as source
2. **Kadrey v. Meta (July 2023)**
    - Authors sued Meta over LLaMA training
    - Similar claims re: Books3
3. **Silverman v. OpenAI (July 2023)**
    - Sarah Silverman, Christopher Golden, Richard Kadrey

Authors' Arguments

Core claims:

- Every book in training data is an infringing copy
- Models can output summaries and passages from books
- Market substitution for book sales and licensing
- Competing use rather than transformative use

Emotional appeal:

> "These are the building blocks of our livelihoods. We should be compensated." - Authors Guild statement

Claimed economic effects:

- "Why buy the book when AI can summarize it?"

Getty Images v. Stability AI

Filed: February 2023 (US and UK)

Plaintiff: Getty Images (stock photo company)

Defendant: Stability AI (creators of Stable Diffusion)

Claims:

- Stable Diffusion trained on 12 million Getty images
- Generated images sometimes include Getty watermark (!)
- Competes with Getty's licensing business
- Violates terms of service

Significance: Focuses on visual works, not text

Music Publishers v. Anthropic

Filed: October 2023

Plaintiffs: Music publishers (Universal, ABKCO, Concord)

Defendant: Anthropic (Claude)

Claims:

- Claude trained on copyrighted song lyrics
- Can reproduce lyrics when prompted
- 500+ songs identified in complaint

Example: Asking Claude for lyrics to "Sweet Caroline" or "American Pie"

Applying Fair Use: Factor 1 (Purpose/Character)

Is AI training transformative?

AI Companies argue YES:

- Training is statistical pattern extraction
- Not replacing the original work's purpose
- New purpose: generating novel outputs
- Like Google Books (search vs. reading)
- Enables new creative tools and research

Creators argue NO:

- Purpose is commercial (selling AI services)
- Outputs can substitute for originals

Is AI Training Transformative? The Debate

Comparison to Google v. Oracle:

- Oracle: Functional interface for compatibility
- AI: Statistical patterns for generation

Comparison to Google Books:

- Google Books showed snippets for search
- AI generates full texts/images that might substitute

Comparison to image search thumbnails:

- Thumbnails: low-res, different purpose (navigation)
- AI images: high-res, same purpose (viewing/using art)

Key question: Does the *purpose* of the copy matter, or the *output*?

Applying Fair Use: Factor 2 (Nature of Work)

What kinds of works?

- **Factual:** News, Wikipedia → favors fair use
- **Creative:** Novels, art, music → disfavors fair use
- **Functional:** Code → favors fair use
- **Published:** Most data → neutral

Problem for AI:

- Training includes BOTH factual and creative works
- Factor 2 cuts both ways

Applying Fair Use: Factor 3 (Amount Used)

How much was copied?

The reality:

- Entire books, articles, images copied for training
- No excerpting or selection
- Wholesale copying of datasets

AI companies' response:

- Entire work needed for effective training
- Individual work contributes minimally to final model
- Like Perfect 10 v. Amazon (entire thumbnails needed)

Creators' response:

Applying Fair Use: Factor 4 (Market Effect)

Does AI training harm the market for copyrighted works?

Creators argue YES:

- AI-generated summaries substitute for buying books
- AI art competes with commissioned artists
- Destroys potential licensing market
- Devalues creative work

AI companies argue NO:

- Different market (tools vs. content)
- May increase demand (exposure effect)
- Most outputs don't substitute for any specific work

The Licensing Market Debate

Creators' position:

- Licensing to AI is a valuable potential market
- Copyright gives them right to control this use
- Many authors would license (for fee)
- OpenAI now pays publishers (AP, Axel Springer) — proves market exists!

AI companies' position:

- Can't retroactively create market harm
- Fair use itself creates transformative markets
- Oracle couldn't claim licensing market either
- Paying publishers now doesn't prove infringement before

The Memorization Problem

What is memorization?

- AI models sometimes reproduce training data verbatim
- More likely with data seen repeatedly
- Can be extracted with specific prompts

Examples:

- GPT-3 reproducing news articles
- Stable Diffusion outputting training images
- Claude reproducing song lyrics (the lawsuit against us!)

Why it matters for fair use:

- Undermines "transformative use" argument

How Much Memorization Happens?

Research findings (Carlini et al.):

- GPT-2 memorized hundreds of examples
- Larger models memorize more
- Training data duplicates increase memorization
- Rare or repeated phrases more likely memorized

Scale question:

- If 0.01% of outputs are memorized, is that acceptable?
- Is ANY memorization copyright infringement?
- Does it matter if extraction requires specific prompts?

Legal question: Is occasional reproduction fatal to fair use defense?

Arguments FOR Fair Use (AI Companies)

1. Transformative purpose

- Learning patterns, not copying expression
- Outputs are new, original works
- Different from reading/viewing original

2. Analogous to human learning

- Artists study other artists
- Writers read books to learn
- Why is AI different?

3. Promotes progress

- Constitutional purpose of copyright

Arguments AGAINST Fair Use (Creators)

1. Commercial exploitation

- Multi-billion dollar companies
- Selling access to models
- Profit from others' work without permission

2. Not truly transformative

- Outputs serve same purpose as originals
- AI art competes with human art
- AI writing competes with human writing

3. Massive scale

- Billions of works copied

Arguments AGAINST Fair Use (Creators/Publishers cont'd)

5. Consent and control

- Creators never agreed
- Can't opt out after the fact
- Loss of creative control

6. Laundry of pirated content

- Books3 from pirate libraries
- Acknowledging illegal sources
- Using infringement to justify infringement?

7. Alternatives exist

The "Human Learning" Analogy

AI companies say: "AI learns like humans learn from examples"

Creators respond:

1. **Scale:** Hundreds of books vs. millions
2. **Purpose:** Personal growth vs. commercial products
3. **Memory:** Imperfect recall vs. verbatim reproduction
4. **Market:** Limited output vs. infinite generation
5. **Law:** First sale doctrine vs. database copying

Question: Is the analogy persuasive or flawed?

International Perspectives: European Union

EU AI Act (2024):

- Transparency about training data required
- Copyright holders can opt-out
- Must publish training data summaries

EU Copyright Directive:

- Text and Data Mining exception
- Research: broad; Commercial: restricted
- Rights holders can reserve rights

Enforcement:

- Italy investigating OpenAI

International Perspectives: United Kingdom

UK approach:

- Proposed broad TDM exception for AI (2022)
- Would allow training without permission
- Strong pushback from creative industries
- Government paused reforms (2023)

Current state:

- Existing TDM exception only for non-commercial research
- Commercial AI training likely requires licensing
- Push for voluntary licensing frameworks

International Perspectives: Japan

Japan's copyright law:

- Broad exception for "information analysis" (since 2018)
- Allows copying for machine learning without permission
- Applies even for commercial use
- Among the most permissive jurisdictions

Rationale:

- Promote AI innovation and competitiveness
- Intermediary copying for analysis
- Outputs still subject to copyright if substantially similar

Result: Many AI companies have operations in Japan

Policy Solutions: Opt-In vs. Opt-Out

Opt-out (AI companies prefer):

- Default: data can be used for training
- Creators can opt-out if desired
- Maximizes training data availability
- Example: robots.txt for web scraping

Opt-in (creators prefer):

- Default: data cannot be used without permission
- Must obtain consent before training
- Respects creator autonomy
- Example: Getty licensing model

Policy Solutions: Licensing Frameworks

Collective licensing:

- Organizations negotiate for creators (like ASCAP/BMI)
- Bulk licensing to AI companies

Voluntary licensing:

- AI companies pay publishers directly
- OpenAI: AP, Axel Springer, Reddit

Compulsory licensing:

- Government-set terms and rates
- Must pay, but can't be blocked

Challenges: Valuation, distribution, research exemptions

Policy Solutions: Technical Measures

Watermarking:

- Embed imperceptible marks in content
- AI training could detect and respect
- Could enable automated opt-out
- Proposed in EU AI Act

Model documentation:

- Require disclosure of training data sources
- "Data nutrition labels"
- Enables rights holders to identify use

Differential privacy in training:

The Innovation vs. Compensation Dilemma

Innovation perspective (tech companies argue):

- Requiring permission would slow development
- Licensing costs would be prohibitive
- Smaller companies couldn't compete
- Public access implications

Creator compensation perspective (creators argue):

- Copyright's incentive structure at stake
- Business model disruption
- Licensing markets exist
- AI companies profit from use

What Do Courts Do When Law Is Unclear?

Judicial considerations:

- Interpret existing law (don't make new policy)
- Consider Congressional intent
- Weigh public interest
- Rely on precedent (Google v. Oracle, Google Books)

Possible outcomes:

1. Fair use applies → AI companies win
2. Fair use doesn't apply → Creators win, seek damages
3. Mixed → Fair use for some uses/data, not others
4. Courts defer to Congress → Ask for legislation

What Might Congress Do?

Possible legislative approaches:

1. **Pro-AI:** Explicitly allow training as fair use or new exception
2. **Pro-creator:** Require licensing, opt-in consent
3. **Compromise:** TDM exception with opt-out + attribution requirements
4. **Sector-specific:** Different rules for research vs. commercial AI

Political challenges:

- Tech industry lobbying (pro-AI)
- Creative industry lobbying (pro-creator)
- Bipartisan interest in AI competitiveness
- Bipartisan interest in protecting creators

Activity: Fair Use Evaluation

Scenario: A startup trains an AI model on 50,000 copyrighted novels to create a "book summary generator" that produces chapter-by-chapter summaries.

Apply the four factors:

1. **Purpose/character:** Transformative? Commercial?
2. **Nature:** Highly creative works?
3. **Amount:** Entire books?
4. **Market effect:** Substitutes for reading? Harms sales?

Your conclusion: Fair use or infringement?

Pair-share (5 min), then discuss as class

Predictions: What Will Happen?

Short term (1-2 years):

- More lawsuits filed
- Some settlements (with NDAs)
- District court rulings (likely split)
- AI companies adopt voluntary licensing for new data

Medium term (3-5 years):

- Circuit court decisions
- Possible Supreme Court case
- Some Congressional hearings, maybe legislation
- Industry standards emerge

Who Benefits from AI?

Current distribution:

- AI companies: revenue, market value
- Users: access to tools
- Creators: uncertain legal status

Different stakeholders want:

- AI companies: clear rules, ability to innovate
- Users: continued access
- Creators: compensation, control

Questions courts will consider: Economic distribution? Incentive structures? Public access?

Policy Questions Beyond Law

Questions raised by stakeholders:

Creators ask:

- Training without consent?
- Business model implications?
- Compensation for contribution?

Tech companies ask:

- Blocking technology development?
- Scope of copyright control?
- Building on existing knowledge?

Courts face: Questions about property rights, economic incentives, innovation policy, and

Key Takeaways

1. **AI training involves copying billions of copyrighted works** at unprecedented scale
2. **Fair use analysis is uncertain** — reasonable arguments on both sides
3. **Memorization weakens transformative use arguments**
4. **Current lawsuits will shape the legal landscape** (NYT, Authors Guild, Getty)
5. **International approaches vary** (Japan permissive, EU restrictive, UK uncertain)
6. **Policy solutions exist** (licensing, opt-in/out, technical measures) but tradeoffs
7. **Balance needed** between innovation and creator compensation

Questions for Discussion

1. Should AI training on copyrighted works be fair use? Why or why not?
2. If you were a judge, how would you apply the four-factor test to AI training?
3. Is the "AI learns like humans" analogy persuasive?
4. Should the law distinguish between non-profit research AI and commercial AI products?
5. How should we compensate creators if their work contributes to valuable AI models?

Looking Forward: Open Questions

Unresolved issues:

- What about AI outputs? Are they copyrightable?
- Who's liable if AI outputs infringe? (User? AI company?)
- How does this affect open-source AI models?
- What about synthetic training data (AI-generated)?
- Can you copyright AI-assisted works?

Rapid evolution:

- Law moves slowly
- Technology moves fast
- Uncertainty will persist

Practical Implications for You

As creators:

- Understand your work may be in training data
- Consider opt-out tools (Glaze, Nightshade for artists)
- Monitor how AI uses your work
- Advocate for desired policies

As AI users:

- Understand legal uncertainty
- Be cautious with commercial AI-generated content
- Consider attribution and ethics
- Stay informed as law develops

Thank You

Questions?

Office hours: [Your schedule]
Email: [Your email]

Next steps:

- Complete copyright assignment
- Follow ongoing lawsuits
- Think about how you'd balance these interests

Additional Resources

Court documents:

- NYT v. OpenAI complaint: https://nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html
- Authors Guild lawsuit: https://authorsguild.org/news/authors-guild-sues-openai-for-systematic-copyright-infringement/

Research:

- Carlini et al., "Extracting Training Data from Large Language Models" (2021)
- Samuelson, "Generative AI Meets Copyright" (2023)

News:

- The Verge AI section: https://www.theverge.com/ai-artificial-intelligence