FAIR USE FOR MACHINE LEARNING IN THE U.S.: LESSONS FROM *WARHOL* AND *GOOGLE BOOKS*

Shyamkrishna Balganesh

Sol Goldman Professor of Law

LLMs/Foundation Models (e.g., ChatGPT)

- ML models that are trained on large amounts of data.
  - Data analysis for predictive purposes in a sequence.
  - Content (data) tokenized into “tokens” for the machine to keep track.
  - Creation of patterns of repetition based on predictive role of tokens.
  - Multiple stages of cleaning of the data, de-duplication, etc.
- Most of the time a local copy of the training data is made.
- The model *ingests* the data and *learns* from it, to be able to produce outputs that may/may not be similar to the training data.
- Is this use of training data copyright infringement?

Type of Data Used

- Where is the data sourced from?
  - Routine reliance on “publicly available”.
  - Much of it is copyright protected, even if publicly accessible.
  - Wrongly equate public accessibility with public domain.
- If the data is entirely public domain.
  - No copyright issues directly implicated in the training.
  - Unless ancillary access-related laws are implicated (e.g., encryption).
- Most successful ML models use publicly available copyright-protected data in significant part (e.g., Stability AI, GPT-4).
- Some rely on repositories with elaborate ToS (e.g., CoPilot/GitHub).

Nature of Copying

- Some countries (e.g., Japan) differentiate based on purpose behind the copying.
    - Copying for non-expressive purposes exempted from infringement.
- Since the ML model is merely tokenizing the data and patterning it, seen as non-expressive.
- U.S. law *does not* contain such a distinction.
    - Once copying exists, its purpose/nature is irrelevant for the prima facie case.
- Instead the non-expressive purpose figures in the fair use analysis.

U.S. Fair Use Provision (17 U.S.C. §107)

Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work…is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—

(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;

(2) the nature of the copyrighted work;

(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and

(4) the effect of the use upon the potential market for or value of the copyrighted work.

The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.

Authors Guild v. Google, Inc. (2015)

- Infringement lawsuit against Google for its famed Google Books project.
    - Scanning of thousands of books.
    - *Search* functionality to locate words; display through the *snippet* function.
    - Local (private) copy of entire books.
- Defendant relied on the fair use doctrine.
    - Primary reliance: transformative use (*Campbell*).
- Judge Pierre Leval runs the analysis through the four fair use factors, and concludes that it was a fair use.

Authors Guild v. Google, Inc. … contd.

- Factor 1:
  - Copying for search was a transformative purpose.
  - Copying was to make available *information about the books* not the books themselves.
  - Purpose very different from the original reading purpose of the books.
  - Snippet view similarly designed to tell searcher *where* a term appears.
  - Commerciality does not outweigh transformativeness.
- Factor 2: Mildly favors fair use since it does not perform a substitutive function.
- Factor 3: Amount and substantiality justified by the transformative purposes.
- Factor 4: No real substitutive market harm that cuts against fair use.

ML Training and *Authors Guild*

- Many mistakenly presume that it *easily* applies.
- Crucial differences that a court will note:
  1. Non-expressive use?
     - *About* the work vs. *the* work itself (enjoyment v. non-enjoyment); when does that breakdown.
     - Enjoyment cannot be purely based on the identity of the actor.
     - Why isn't ingestion (tokenization) more like *translation*?
  2. Commerciality?
     - Court assumes it is secondary; but *Warhol* changes that.
  3. Market Effect
     - Potentially significant; cannot be wished away.

AWF v. Goldsmith (2023)

- Straightforward facts; yet reopening the transformative use debate.
- Focus entirely on factor one; no transformative purpose.
- Lessons:
    1. Transformativeness is a matter of degree; to be balanced against commerciality.
    2. New meaning or message cannot be considered in isolation.
    3. No such thing as a transformative work; fair use focuses on the “use”.
    4. Each use is the unit of analysis for fair use.
    5. Purpose to be:
        - Narrowly calibrated by looking to the market and substitutive effect;
        - Understood objectively (“reasonably be perceived”).
        - Balanced against the derivative works right
- Citations to *Authors Guild* not validation of the outcome there.

AWF (Warhol) and ML Training

- Focus should be on the individual use, not the resulting work.
- Purpose to be calibrated carefully.
    - "Non-expressive" unlikely to be a satisfactory category.
    - *Tokenization* and *disaggregation*: are these transformative?
        - Potentially within the scope of the derivative works right?
        - Is tokenization the creation of a derivative? Is it a translation?
    - Commerciality will loom large.
- The output question will re-emerge under factor four.

Beyond Fair Use: *Doe v. GitHub*

- GitHub: well-known online repository that hosts software source code. Much of it is public and OS.
    - Uploaded code subject to GitHub's expansive ToS.
    - When OS code is uploaded, subject to terms of OS licenses.
- GitHub develops Co-Pilot, an AI-base coding assistant, that suggests code to developers who subscribe to it.
- Co-Pilot employs ML to train the model, and uses/reproduces code uploaded to GitHub.
    - ML disregards attribution, copyright notices and license terms.
- Potential class action against GitHub for several causes: DMCA, breach of licenses, several state law claims.
    - DMCA claim for removal of CMI rather than infringement because of OS license.

Doe v. GitHub … latest ruling (May 11, 2023)

- Standing found, but limited:
  - Must be specific; not enough for privacy.
  - For property rights – particularization needed; specific connection.
  - Own code specifically used in output – needed, not shown.
  - Future harm conferred sufficient standing, but only for an injunction.
- Allowed to proceed, but some claims eliminated (Motion to Dismiss).
  - DMCA 1202(b) claim allowed to proceed: removal of CMI.
  - Breach of OS license terms also allowed to proceed.

Should we rely on Fair Use?

- Case-by-case and fact-specific – not ideal.
    - Each ML model varies and has significant differences.
- Alternatives?
    - Legislative collective licensing? (operational nightmare)
    - Categorical exemption (one-sided)
    - Categorical liability (one-sided)
    - Negotiated compromise needed – question is where to set the default.
- Crucial to understand and appreciate the technology underlying the models and variations.