Copyright:
Right Answer for Open Source Code,
Wrong Answer for Open Source AI?

McCoy Smith
Lex Pan Law & Opsequio
Portland, Oregon, USA

Massive (Non-Legal) Disclaimer

- AI is complex, there are a lot of different techniques used
- © and AI is a huge matter of current legal dispute (both in courts, and in government), and a lot of the issues are not yet clearly resolved
- This presentation abstracts a lot, both on the technical side, and on the legal side

Open Source

- Both:
  1. A philosophy
     - Software should be “free” (“as in speech”)
  2. A way of defining how legal rights are conveyed
     - The Open Source Definition is primarily #2
       - Free Software Definition more of a hybrid
       - OSI license review decisions sometimes consider #1, but primarily focus on #2

Defining Legal Rights

- The Free and Open Source movement has primarily/exclusively focused on using © as the legal mechanism to underpin its philosophy
  - "Copyleft"
    - Using a proprietary legal regime (©) to achieve an anti-proprietary result
  - "Permissive"/"Non-copyleft"
    - Earliest versions were styled as primarily directed to conveying © rights
      - Although later versions also cover (P), © is still the main mechanism

© As A Legal Regime

- © has a long history of protecting “creative” works
  - Literature
  - Art
  - Music
- © protecting software & technology a fairly recent development
  - US Congress in 1970s examined question of © and software
    - Ultimately concluded that software should fall within ©
      - There was a minority who argued software primarily or exclusively “functional” and therefore not © “creative”
  - Other countries followed US lead

The Limits of © For Technology

- © only covers “expression” not “ideas”
  - (P) is designed to cover ideas
  - The longstanding, and almost never answered, question:
    - What is “expressive” in software?

The Limits of © For Technology

- Anti-©, pro free “speech,” exceptions to ©:
  - Merger:
    - If expressive, but that expression is the only way to achieve a function, no ©
    - “Merger” is a US law concept, but a similar analysis applies outside US
  - “Fair use”/“Fair dealing”
    - A defense: even when there is © infringement, certain infringements are acceptable
      - US (“fair use”) vs ROW (“fair dealing”) have significantly different algorithms for analyzing this question
        - Fair use has become a default mechanism in many controversial cases in US
          - Examples: Oracle vs Google (API case); Author’s Guild v. Google (Google Books case)

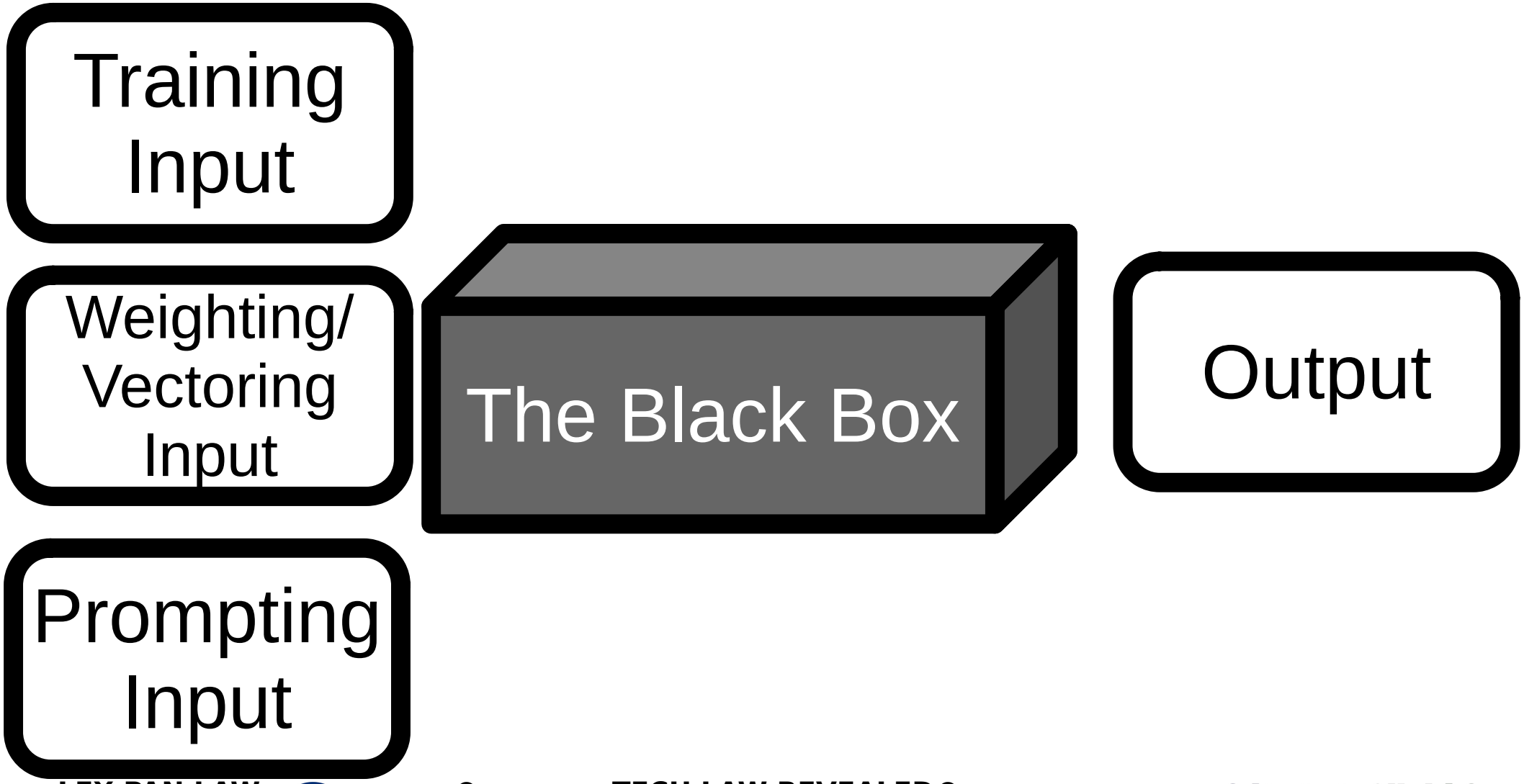
The Limits of © For Technology

- Data
  - Data, per se, is not subject to ©
  - Databases (i.e., large compilations of data) might be ©
    - Depending on where you are
      - US: in almost all circumstances, not ©
        - Feist v. Rural Telephone Service, 499 U.S. 340 (1991)
          - Unless there are “creative choices” in putting the data together
      - EU: European database directive allows ©
        - For “a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means”
        - But does permit “extracting and re-utilizing insubstantial parts” of database without being an infringer

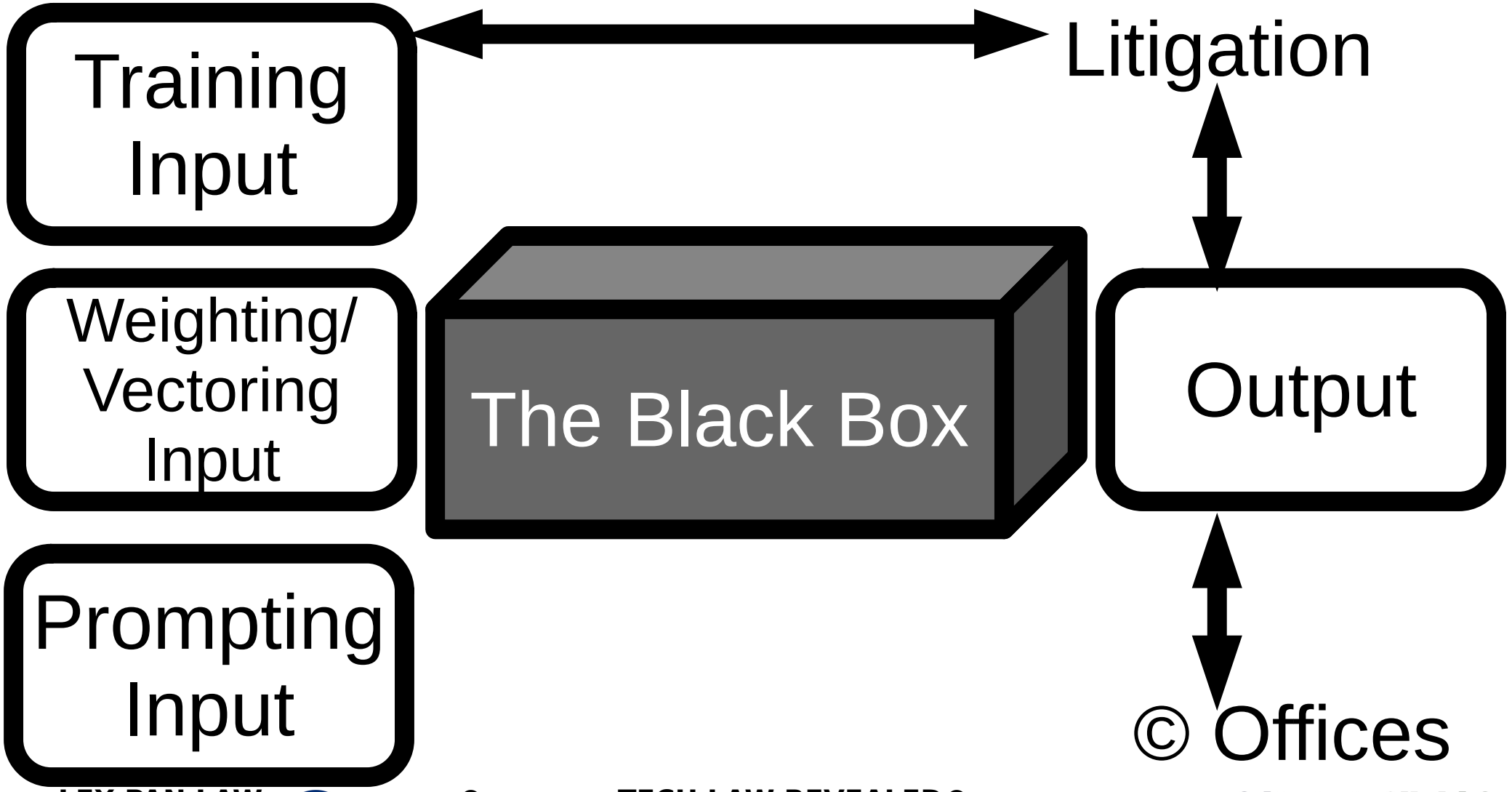
Open Source © Licensing for Software

- The model has worked, reasonably well, for 40+ years
- The speedbumps:
  - License incompatibility
    - Generally a copyleft ⟷ copyleft issue
    - Sometimes a copyleft → non-copyleft issue
  - Attribution & license notice requirements
    - Generally not a copyleft issue
    - Possibly an issue for non-copyleft/permissive
      - If there is a mixture of many different non-copyleft licenses
      - Or a huge number of © notices

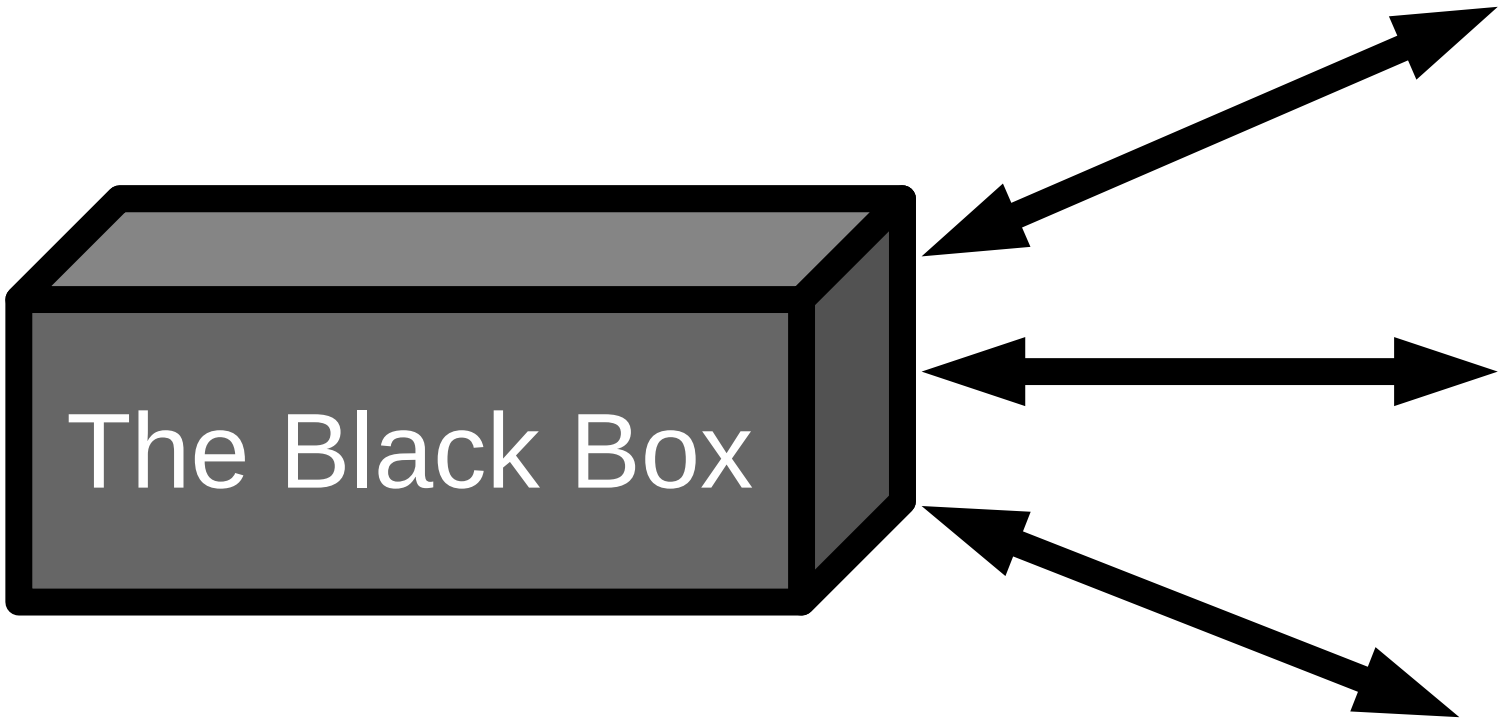
Open © Licensing for AI



Open © Licensing for AI



What's In the Black Box?

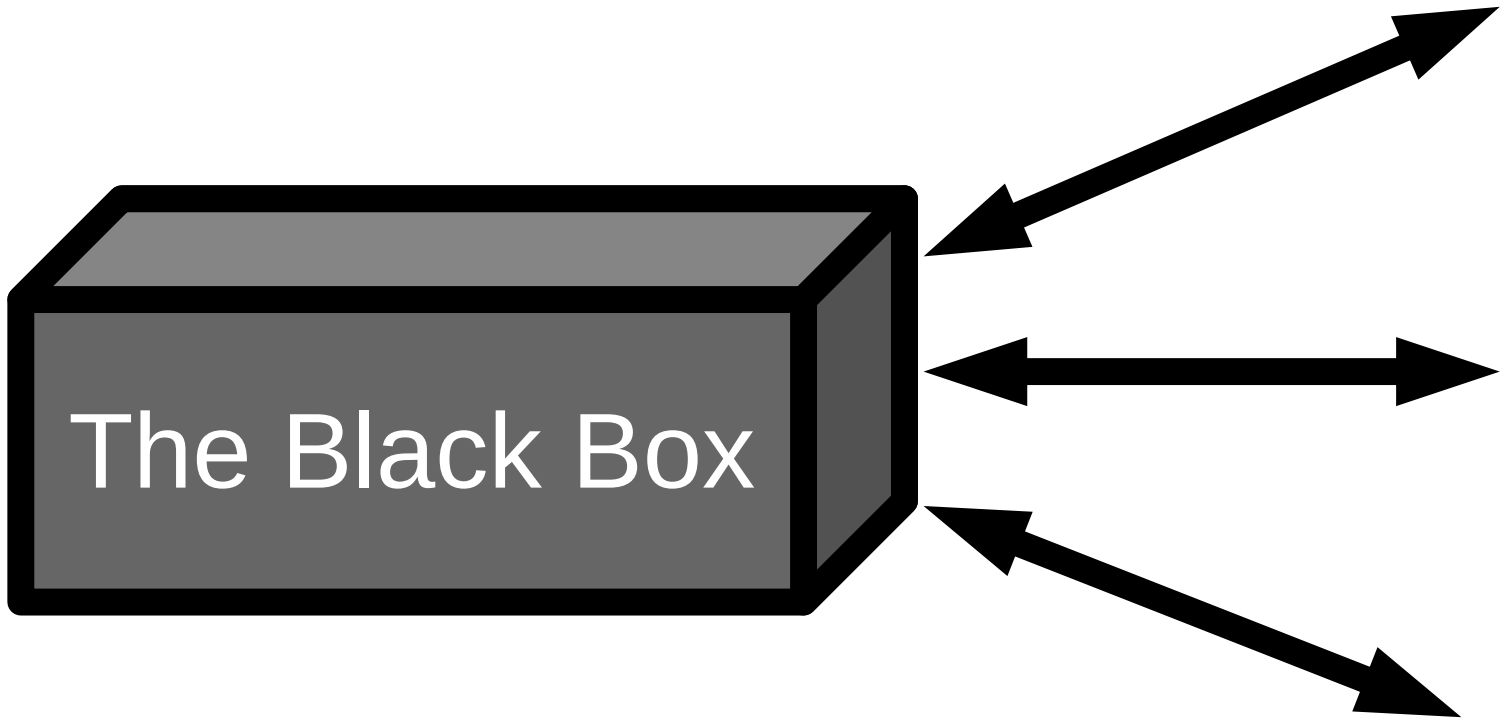


The weights/vectors, created by applying the base logic to training data, alone or with weighting/vectoring input

Copies of the training data, fully intact, partially intact, or fully abstracted from any expression

The base logic for how to handle incoming training data and creating weights/vectors from it, and using those values to produce output

What's In the Black Box?



The weights/vectors, created by applying the base logic to training data, alone or with weighting/vectoring input

Not © (except database rights)? May not be licenseable

Copies of the training data, fully intact, partially intact, or fully abstracted from any expression

© (?) Comes with existing license (?)

The base logic for how to handle incoming training data and creating weights/vectors from it, and using those values to produce output

© Open Source Licenseable

What Is In the Black Box?

- A collage-maker
  - An assembler of a semi-coherent concatenation of bits and pieces of © content
- A real general intelligence
  - An extractor of the vocabulary, syntax, grammar, and other rules from which © content is built – including by humans – and applier of those rules to create new expression
- Something in-between
- Does the answer depend on how big is the training dataset, for how long the training happens, and how far output is removed from the human-input part of the weighting/vectoring?

Freedom Zero

- “The freedom to run the program as you wish, for any purpose”
  - Generally, interpreted to mean your private use, copying, etc. not subject to FOSS obligations
    - Philosophical rationale: you have the right to keep private your private modifications
    - Practical rationale: having to open up your private modifications becomes a compliance headache
      - How often, and to whom?

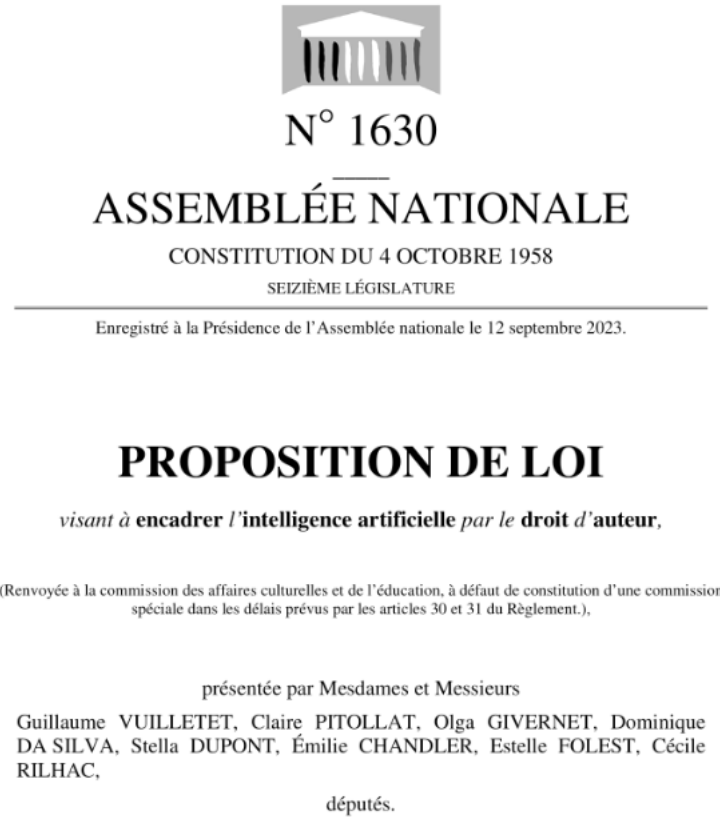
Freedom Zero

- Copying © code into memory, for training, without output of any © code: an exercise of Freedom Zero?
- If your weights/vectors stay internal (are not part of any output), wouldn't Freedom Zero govern them?
  - And if not, how does one make "open" these parameters which are massively, and constantly, in a state of flux?

Idle Speculation & Prediction

- Whether or not AI is now engaged in any sort of © infringement – even if it is only temporary internal storage of content for the process of training – AI likely to evolve to the point where the case that it is using © is tenuous or non-existent
  - And this may not be a fair use/fair dealing analysis
    - Meaning, it's not a multi-factored factual inquiry
  - Functionality & merger, idea vs expression, may end up being the governing questions
    - Also factual inquiry, but much less amenable to "balancing"
- If so, the paradigm that has been used to ensure openness in Open Source for 40+ years – ©, and licensing – may no longer work
- What then, is to be done?

Is Legislation The Answer?



N° 1630

ASSEMBLÉE NATIONALE

CONSTITUTION DU 4 OCTOBRE 1958

SEIZIÈME LÉGISLATURE

Enregistré à la Présidence de l'Assemblée nationale le 12 septembre 2023.

PROPOSITION DE LOI

visant à **encadrer** *l'***intelligence artificielle** *par le* **droit** *d'***auteur**,

(Renvoyée à la commission des affaires culturelles et de l'éducation, à défaut de constitution d'une commission spéciale dans les délais prévus par les articles 30 et 31 du Règlement.),

présentée par Mesdames et Messieurs

Guillaume VUILLETET, Claire PITOLLAT, Olga GIVERNET, Dominique DA SILVA, Stella DUPONT, Émilie CHANDLER, Estelle FOLEST, Cécile RILHAC,

députés.

- Appears to mainly deal with the © input issue
- Other legislative initiatives (US) focused on regulation and safety, not legal protections

Is Legislation The Answer?

- Something like the EU database directive, but specifically encompassing machine generated and refined weights/vectors?
  - Might be hard to get enacted in US
    - © for non-human created things potentially implicates U.S. Constitution
  - A double-edged sword
    - By providing proprietary rights to weighting/vectoring, allows easier and more aggressive proprietization of AI training by those entities that don't want to be open
    - But, would allow a weight/vector creator to use an open source license to limit or control proprietization downstream
      - Is this only really an issue for copyleft?
      - If weights are essentially public domain, by releasing you've given permission to use without restriction

Is Contract The Answer?

- Open Source by contract has always been a subject of debate
  - Some licenses specifically state they are enforceable contracts
  - Others are silent, or might even say (indirectly) the are only enforceable via ©
- Contract introduces its own set of complications in a international, multi-party, Open Source development environment
  - Many hundreds of different rules for interpreting a contract and its effect on downstream behavior
  - Some places, like US, may have to worry about © law overriding (preempting) contract law

Are Public Pledges The Answer?

- Pledges have some history in Open Source as mechanism to signal what is good behavior
  - Various pledges not to enforce patents against Open Source in the ‘00s
  - GPL Cooperation Commitment
    - Pledging to not enforce GPL in a potentially harsh and harmful way
- The efficacy of pledges depends upon:
  - Their widespread adoption
  - Their effect on behavior of others
  - Their enforceability/Resistance to being later retracted

Time for a Rethink

- Lots of presenters talking about, philosophically, what is “open” and AI
  - Very likely will result in a rethink of how we think of “open”
- Very likely also a need to talk about, legally, how to get to the philosophical ideal of “open” and AI
  - And doing so may require:
    - Upending of some of the fundamental legal premises of the past 40+ years
    - Advocacy for legislative changes
    - Widespread consensus among AI creators of rules to follow

Further Reading

Further Reading

- Many academics looking at issue of © & AI
  - Not focused on AI & “open”
- “A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs” Dr. Andres Guadamuz (U. Sussex)
  - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4371204
- “Talkin’‘Bout AI Generation: Copyright and the Generative-AI Supply Chain” Katherine Lee, A. Feder Cooper, James Grimmelmann (Cornell U.)
  - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4523551
- “How Generative AI Turns Copyright Upside Down” Mark A. Lemley (Stanford U.)
  - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4517702
- “Foundation Models and Fair Use” Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, Percy Liang
  - https://arxiv.org/pdf/2303.15715.pdf