Law 758 – Emerging Law of Artificial Intelligence Outline

Brad N Greenwood

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war. – *Center for AI Safety*

Definitions and General Background

- Definitions (Abridged)
    - Agent: system or program that is capable of autonomously performing action
    - Algorithm: logical sequence of steps written in code to perform a task
    - AI: Artificial Intelligence - Machine based system that can, for a given set of defined objectives, make predictions
    - AGI: Artificial General Intelligence - type of AI that matches or surpasses human cognitive capabilities across a wide range of cognitive tasks. Typically, not trained on a specific task
    - CBRN – Chemical, Biological, Radiological, and Nuclear Weapons
    - CSAM: Child Sexual Abuse Material
    - Deep fake: synthetic image that is manipulated or wholly generated by AI
    - Discrimination (warfare) - Combatants must be distinguished from civilians, and attacks must be directed only at legitimate military targets.
    - Dual Use Foundation Models - AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters. See National Security section
    - Fair use: fair use is any copying of copyrighted material done for a limited and "transformative" purpose, such as to comment upon, criticize, or parody
    - Foundational model: AI model that is trained on massive and diverse datasets and can be applied across a wide variety of use cases. See Dual Use Foundation Model
    - Generative AI: a category AI that can generate new output based on training data
        - Chat GPT (Chat [based] Generative Pretrained Transformer)
    - Hallucination: occurs when a model generates inaccurate, misleading, or nonsensical information while presenting it as if it were factual
    - Humans and Loops:
        - Human-in-the-loop (HITL) refers to a method where human intelligence and AI systems collaborate to achieve a desired outcome
        - Human-on-the-loop (HOTL) involves humans providing oversight and correcting results after an AI system has produced them
        - Full Automation means no human involvement. Weapons free.
    - Jailbreak: prompt that induces a tool to bypass rules and restrictions
    - Liars Dividend: as people learn that deep fakes are increasingly realistic, false claims that real content is AI-generated become more persuasive too.
    - Machine learning: iterative process whereby a machine gets better at a task, as defined by a performance metric, as it performs that task more (i.e. becomes more experienced)
    - Model: software configuration (algorithm) that can make inferences once fed new data
    - Model weights: numerical value that amplifies or suppresses a pattern found in data
        - Think of this as how much some factor affects classification into bucket a v bucket b. Ears and nose for dogs and cats. Or just think of it as a beta coefficient
    - NCII – term of art – Non-consensual intimate images

    - Open-source model: models and software that are publicly available and can be modified or manipulated by anyone
        - Closed source: opposite. Code is not publicly available. Cannot be manipulated by the public. Parallel to Windows OS (closed) v Unix OS (open)
    - Predictive AI: category of AI that uses previous data to forecast future data
        - Not necessarily temporal (future) data, just new data to the system
    - Red-Teaming: ethical hacking. Simulates real-world cyberattacks to identify vulnerabilities and weaknesses. Helps test security defenses, improve incident response, and enhance their overall security posture
    - Structure v Unstructured data: you know this. Format and organization
        - Structured data fit neatly into spreadsheets
        - Unstructured data have no predefined format (e.g. social media, images)
    - Transformers: neural nets that learn context and understanding through sequential data analysis. They use modern mathematical techniques to identify how data elements influence and depend on each other
- What is AI?
    - Two types, Artificial General Intelligence (AGI) and Directed AI
        - Intelligence in the general sense (AGI), reaction to environment. General Intelligence implies that the system will be able to execute on tasks that it is not specifically trained on.
        - Directed AI – solution to simple process
    - Both are algorithmic processes for making predictions. The AI above are subsets of the general phenomenon
- What is artificial general intelligence (AGI)?
    - More specific than above.
    - General intelligence is an AI that mimics human learning. It learns generally, and not necessarily for a prescribed and predefined task. The goal is for this to be at or better than an average human. Super intelligence is the goal
        - This is different from Directed AI because Directed AIs solving a specific problem (chihuahua vs blueberry muffin)
        - AGI is instead about extrapolating answers it is not focally trained on
    - Super intelligence is broadly defined as the system being significantly smarter / better at learning than the average human
- What are the components of an AI system?
    - Three required components – Data (for training), Algorithm (software), Hardware (Chips)
    - Data (training) – these are the component pieces the algorithm learns on to make its predictions
    - Hardware – GPUs (chips) – these chips are designed for parallel processing which can do multiple calculations at the same time. GPUs are scaled by operations in parallel
        - NVIDIA and AMD are the largest manufacturers. NVIDIA by far the largest
    - Algorithm (software) – typically we are dealing with neural network training
        - An algorithm is a process or set of rules followed by an information system when executing its operations or solving a problem
        - In a standard AI like Chat GPT, the algorithm represents things in a vector space. The vectors of numbers represent an abstraction of the item you are modeling
            - A housecat is closer to a lion than an apple in vector space
            - There are 12k+ tokens for each item abstracted in GPT-3

        - The algorithm gives us a prediction of what comes "next" for the LLM
        - The prediction is not necessarily giving you the BEST prediction for the next word, it is the BEST prediction as weighted by some element of randomness (temperature)
            - For AlphaGo or Deep Blue - seemingly suboptimal move that is a local non-maxima but ends up being a global maxima
            - For ChatGPT, the temperature permits the answers to change slightly because it does not always choose the global maxima
- How are LLMs and Other AI systems Trained?
    - For a Gen AI, the world wide web is usually scraped. This includes repositories like Common Crawl, WebText, WebText2, Books1, Books2, Wikipedia, and other sites
    - The data is then converted to tokens and the machine learning algorithm is exposed to the tokenized data
    - Model makes a guess (classification or a prediction) which is then evaluated
        - This initial guess is usually hot garbage. The weighting and reweighting improves
    - Weights and re-weights are applied to the various parameters to improve performance. The weighting and re-weighting occurs recursively, leaving the model pre-trained
        - Performance is evaluated against a holdout sample which is not contained within the corpus. Performance can also be evaluated by a human or other agent. This could be through selection of superior output options or some other metric
    - The model is then fine-tuned, potentially by hand, and put into the hosting environment
        - Example: NY Times style fine tuning by Open AI
- What are the common repositories?
    - Common Crawl - non-profit organization that maintains open repository of web crawl data
    - WebText - internal OpenAI corpus created by scraping web pages based on document quality. It includes all outbound links from Reddit which received at least 3 upvotes
    - WebText2 - enhanced corpus of the original WebText covering all Reddit submissions from 2005 up until April 2020
    - Books1 + Books2 – contents are a bit ambiguous. Broadly, these are two internet-based books corpora, containing a large (potentially random) sampling of public domain books
    - Wikipedia – Online user generated encyclopedia
- How big are these models?
    - Deep Seek R1: uses 671 B parameters with 37B parameters per query. Focus is on efficiency
        - Deep Seek's whole thing is a "mixture of experts" approach so not all nodes are used every time
        - Open Source Model
    - GPT-4: 170 billion parameters. ChatGPT uses a dense transformer architecture where all parameters are active for every task
        - Closed source model
    - FERRARO is skeptical of the DeepSeek claims (PRC) based on when (timing) it was made available to the public globally
- Are AI Systems Intelligent?
    - Superintelligence is a waiting game. This is the received wisdom, but if not superintelligence, then the ability to perform tasks a human might do much faster than them (e.g. classification)
    - Contrast – Stochastic Parrot Theory – This theory argues that the AI is just learning patterns and parroting them back to the user. It is not inventing a new pattern, it is just a very clever monkey typing Shakespeare with no semantic understanding

        - JAFFER argues that this is a distinction without a difference because we are all stochastic parrots. What is it to be human? Is there free will? What is it to be Youngjin?
- For Govt - Core Policy Issues Surrounding AI (Abridged List)?
    - National Security Issues
        - Safeguarding high performing algorithms (both the perspective of withholding them from hostile entities and maintaining global technological dominance)
        - Ability to affect Chemical, Biological, Radiological, and Nuclear Weapons (CBRN)
        - Cyber security (both offensive and defensive)
        - Export controls / classification – export controls restrict the movement of intellectual capital (e.g. algorithms, weights) and physical capital (e.g. chips) to maintain national interests
        - Governance – Should there be an AI Agency? Who would lead it? Should the specialized agencies that have subject matter expertise deal with the individual concerns that manifest from AI.
            - Consider this a DNI debate. Coordinating agency.
    - Open v Closed Source Systems – Balancing Issue. Open source leaves the potential for harmful skills to fall into the hands of bad actors. Closed source systems stymie the ability for agents to collaborate (cross fertilization of ideas).
        - ChatGPT is a closed-source AI model developed by OpenAI
        - TensorFlow and PyTorch are open-source AI libraries for deep learning
    - Talent and Immigration
        - How do policy makers ensure we train the public for effective and safe use?
        - Technical and nontechnical education and upskilling for use?
        - How do we ensure global talent wishes to do their work in the US?
        - How do we ensure a brain train problem doesn't prevent specialized knowledge from exiting the US (intellectual capital export control problem)?
    - Incentivizing Research
        - Development of effective SOFTWARE and HARDWARE?
        - Public private partnerships? Academia? Grant based research houses (e.g. RAND)?
    - Licensing for operation and use
        - Based on: Size of firm? Size of the model? What the model is used for? Who is using the model?
            - This is a canonical regulation of the THING v USE issue.
    - Intellectual Property Issues
        - These issues emerge at the training phase (data used), output phase (has the output infringed), and the model itself (e.g. stolen weights)
        - Can the AI model own material?
        - Can the model generate IP?
        - Can the model infringe on IP?
        - Who is liable for any infringement that occurs?
    - Privacy (data privacy, PII, training data)
        - Is the AI inferring or training on protected data?
        - Are users feeding protected information into the AI that it is subsequently incorporating into future decision making?
    - Constitutional issues – bias in outputs that disfavors protected classes
    - Environmental Issues – Training the models requires massive infrastructure (physical plant to house hardware and electricity consumption)

    - Supply chain control – relates to the security of both the hardware (chips), software (algorithm), and human capital (researchers) which constitute supply chain
        - Can apply to any component of the AI: Chips, Algo, Data
    - Speech, fair use, content moderation
    - Unemployment, labor displacement (tasks v jobs distinction remains important for this)
    - Crimes and Torts – intent. Commission of a crime by the AI or commission when using the AI in order to aid an unlawful endeavor
    - Deepfakes and sexual imagery
        - Extortion using NCII, fake or real
    - Facial recognition
    - Need to understand how and why AI tech works – models are frequently black boxed and weights cannot be recovered without significant effort.
        - Given the number of parameters that are modeled, and the level of abstraction when creating primitives, it can be difficult or even impossible to explain why X is the outcome of Y test.
        - Explainability – do we know why the AI arrives at its decision
    - Consumers are worried about PII being collected. Is that collection lawful?
    - Disinformation, inaccuracies, and hallucinations
        - Couples with Liar's Dividend issues
    - Governance requirements for the development and deployment of AI
    - Raising questions about ownership and rights

- For Individuals - AI Risks (abridged list)
    - Dignity: Usurping human dignity. Machines performing tasks or harming a person
        - Displacement: This is a jobs and tasks issue. JAFFER and FERRARO focus a lot on automation. I worry the tasks / jobs distinction has gotten lost here
            - Speculative issue remains whether the resultant corpus of tasks allows the person to i) upskill, ii) do more work, iii) focus on value added tasks; or whether true displacement occurs.
            - We might employ fewer associates in a law firm because there is simply less doc review to perform but the same body of overall work. We might employ more associates at the firm because the associates are more efficient and we can take on more work
                - Burnout argument? Same number of associates but less stress is put on them to accomplish the same amount of work.
        - Bias Against Groups: This could be a negative or positive. This could be a harm (facial recognition being demonstrably worse for racial minorities) or improving outcomes for a protected class (the AI is able to mitigate systemic bias in human decision making)
        - Sextortion: Extorting a person by threatening to reveal evidence, or falsely generated evidence, of their sexual activity. Can also encompass sexual manipulation (e.g. chat bot convinces a person to send sexualized images)
            - Obvious issue is deep fake pornography, but this also includes attempting to trick service members into betraying the US govt using fake Instagram models. Manti Te'o issue.
    - Deception: People not understanding what they are interacting with
        - Mimicry: (AI anthropomorphizing) – The AI acting like humans when it is not

        - Forgery and Impersonation (fake information is perceived as real): AI that misrepresents reality and makes people think they are seeing real things that are actually fake. Defrauding people. Political misdirection
        - Liars Dividend (real information is perceived as fake): Opposite of forgery. Makes people think the REAL is FAKE, rather than the forgery which makes people think the FAKE is REAL.
    - Dollars:
        - Compensation: who gets paid, how much, and for what (training data, tuning, image, likeness, intellectual property)
        - Liability: how does replacing human judgement, and who is liable
        - Pricing: how do we price business opportunities
- For Firms - Legal and Business Risks (Abridged List)
    - Discrimination Risks – hiring and promotion process. Producing unfair or discriminatory outcomes based on race / gender / protected class. Federal Regulators (FTC, DOJ, CFBP, EEOC), White House Blueprint, and others highlight this issue. High-risk EU issue.
        - Algorithmic bias in ML hiring, e.g. Amazon hiring software (abandoned)
    - Disinformation Risks - Creation of false information at scale. False narratives.
        - In May 2023, photos (fake) showed a Pentagon explosion and markets panicked.
        - Others: Audio deepfake of Biden telling voters not to vote in NH. *League of Women Voters of New Hampshire v Kramer. In the Matter of Steven Kramer*
        - Text targeting of Black voters to "Text vote for Hillary". Conspiracy case. Eastern district of NY. *United States v. Mackey*
        - Malicious actors can feed bad information into the AI in order to induce hallucination
    - Data Privacy
        - Data as input – are firms permitted to use data when training the LLM?
            - Potential licensing issue. *NY Times v Open AI*
        - Specific Used Cases – consent and PII (training on images of your children without consent, consent to use health information)
        - Access rights of data subjects – Do individuals have the right to know what information is used? Right to be forgotten (EU and California)
        - Leakage – what happens when that that is used in training becomes publicly facing? Is the AI licensed to produce this information publicly? Is it violating copyright? Is the AI accurately representing the data if it is revealed publicly? *NY Times v Open AI*
        - Prompt ingestion – prompt that includes PII or other protected information gets ingested into the algorithm and then into the algorithm's training data, potentially exposing it to others.
            - This could be internal to the AI operator or to the public
            - Information could be protected contractually or other legal means
        - Does the use of the data constitute fair use? Fair use is any copying (reproduction) of copyrighted material done for a limited and "transformative" purpose, such as to comment upon, criticize, or parody a copyrighted work. Such copying can be done without permission from the copyright owner
            - More below
    - IP Risks
        - Inputs: the training data frequently used protected information, trademarks, etc. for which authorization has not be obtained in advance
        - Outputs: Who owns the output of the model or any resulting intellectual property.

        - What if the generated content violates copyright? (e.g. Midjourney making the Simpsons)
        - Is the owner the operator of the AI or the user the operator of the AI? Much of this will be managed by terms of service (TOS)
    - Only AI Generation: US Copyright Office currently indicates that content exclusively generated by an AI absent post prompt manipulation cannot be copyrighted. *Zarya of the Dawn*
        - This is in contrast to AI in a film where there is significant post production. That post production work on the AI generated content permits the copyright of the firm.
        - Recall that in *Zarya*, the characters and the story were still copyrighted, just not the pictures
- Deceptive Trade Practices Risk
    - Companies outsourcing work to a chatbot or AI software when a consumer believes he or she is dealing with a human and marketing it as "human made"
        - This is a base disclosure issue which will be governed by statute or contract
        - Goes both ways. Frequently a firm cannot say its AI if it's a human
    - Both the FTC and WH have emphasized that Section 5 of the FTC Act covers deceiving customers in this way
        - The FTC has taken action against adultery themed chatbots marketing themselves as actual women.
        - Generating fake online reviews. *In the Matter of Rytr LLC.*
        - FTC complaint against RoboLawyer firm to assist consumers with civil issues (e.g. claiming rebates, cancelling subscriptions). *In the Matter of DONOTPAY, INC.*
    - Defective AI systems where you can easily jailbreak the system to produce harmful content. Children are an issue here. *AF v Character AI. Garcia v Character AI.*
- Contract Risks
    - Prompts are not necessarily protected, what you put out IS NOT PRIVATE
    - Business usually put confidentiality controls into contracts
    - Don't use AI with a contract if it has these provisions (or using an AI if you don't say you are using an AI)
    - Inputs and outputs could be discoverable in litigation
- Ethical Risks
    - Companies and people are regulated by professional ethics organizations (lawyers, physicians)
        - State law - Colorado
    - May 2023 - *Mata v Aviana,* plaintiff filed a motion to dismiss because the complaint was filed using made up cases
- Cybersecurity Risks
    - AI would be leveraged to developed malware used in cyberattacks. Adversarial governments are obviously and issue here (Homeland Threat Assessment 2024)
    - AI systems can be a target
    - AI can be used to attack
    - AI can be used to improve security to identify vulnerabilities and respond to threats
        - This is an active and ongoing issue in practice on all fronts
        - Defenders dilemma

- Government Contract Risks
    - US Government is the largest purchaser in the world. There are incredibly complex rules and they impose a lot of obligations
    - When you are preparing a bid, you need to be transparent about use (False Claims)
    - You could get similar bids from similar prompts. This could violate the Procurement Integrity Act. *41 U.S.C. § 2101–2107*
    - AI might train on the input
    - Standard contract issues, does the contract permit the use of AI
- Validation Risks
    - Hallucination issues (AI gets it wrong)
    - Are we overly reliant on the content that comes out of it, is it valid
    - Has a malicious actor induced the AI to hallucinate?

Questions to Ask

- Who is the party we are dealing with?
    - Is it a government? Think about the policy stuff above
    - Is it an individual? Consider the AI Risks above
    - Is it a firm? Consider the Legal and Business Risks above

- What is the underpinning legal issue?
    - Are we dealing with an invention (patent)?
    - Are we dealing with the creation of a copyright?
    - Are we dealing with infringement?

- Civil Procedure: Are we in motion to dismiss or summary judgement?
    - A motion to dismiss challenges the legal sufficiency of a claim - seeks to dismiss a lawsuit before trial based on legal issues like jurisdiction or the failure to state a claim
        - 12(b)6 motion. See below
    - A motion for summary judgment challenges the factual basis of the case - seeks to resolve a case without a trial by arguing there are no genuine disputes of material fact.
        - Sustained summary judgement would prevent defendant from using a particular defense or plaintiff from proceeding with a theory of harm

- Does plaintiff have standing?
    - To establish standing, a plaintiff must demonstrate: 1) a concrete and particularized injury, 2) that the injury is fairly traceable to the defendant's actions, and 3) that the injury is likely to be redressed by a favorable court decision. *Lujan v Defenders of Wildlife*

- What is the level of scrutiny need we concern ourselves with?
    - If we are dealing with patent invalidation, that is an APA issue. Arbitrary and Capricious
    - If we are motion to dismiss? To survive a motion to dismiss (12(b)6 motion)
        - Subject matter and personal jurisdiction (don't assume or state if assuming)
        - The court must be able to provide adequate remedy
        - Complaint must contain sufficient factual matter, accepted as true, to 'state a claim to relief that is plausible on its face.'" *Ashcroft v. Iqbal, Bell Atl. Corp. v. Twombly.*
        - We assume all well-pleaded factual allegations to be true, and determine whether they plausibly give rise to an entitlement to relief. *Faber*
        - Threadbare recitals of the elements are with conclusions are insufficient. *Iqbal*

        - The claims as plead must have at least a plausible chance of success. *Levitt v. Yelp! Inc*

Intellectual Property Production (Patents)

- Statutory Authority for Patenting?
    - Constitution - Article I, Section 8, Clause 8 – "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."
    - The Patent Act of 1952. *35 U.S.C.*.
- Does the invention meet the requirements for a patent? There are five elements of a valid patent
    - Patentable subject matter - any process, machine, manufacture, composition of matter or improvement of matter.
        - Laws of nature, physical phenomena and abstract ideas are not patentable subject matters. *Diamond v. Diehr*
    - Utility - inventions must be useful. This means the item being patented has a credible, specific and substantial purpose. *Patent Act*
    - Novelty – Two requirements
        - Novelty bar - the invention was not known or used by others (i.e., new)
        - Statutory bar - the patented item must not have already been in public use or for sale in the U.S. for more than one year prior to the date the patent was applied for.
    - Nonobviousness - A patent claim is nonobvious if the improvement goes beyond the predictable use of prior art according to its established functions.
    - Enablement - Patent application must include a written description of the item being patented, including the manner and process of making and using it. It must be sufficiently clear that those with ordinary skill in the art would be able to reproduce and use the item
        - Denial of a patent by the USPTO can be appealed pursuant to the Administrative Procedures Act. In the 4th Circuit these appeals are reviewed *de novo*. *Gallagher v. Reliance Standard Life Ins. Co*
        - STANDARD OF REVIEW: Arbitrary and capricious application of the law (*APA*)
- Test for Subject Matter Eligibility of AI Patents
    - Step 1 - examine whether the claimed invention falls within one of the four statutory categories: processes, machines, manufactures, or compositions of matter.
        - For AI innovations, they usually qualify as processes or machines.
        - If successful, go to Step 2A
    - Step 2A - examine whether the claim recites a judicial exception such as an abstract idea
        - Step 2A, Prong One: Does the AI Invention Recite an Abstract Idea?
            - When evaluating AI patent claims, the USPTO distinguishes between claims that "recite" an abstract idea (requiring further eligibility analysis) and those that merely involve or are based on an abstract idea.
                - "reciting an abstract idea" means that a claim (a specific description of the invention) mentions a concept or idea that is considered fundamental and not patentable on its own
            - Three categories of abstract ideas are particularly relevant to AI: mathematical concepts, certain methods of organizing human activity, and mental processes
            - If the claim merely recites an abstract idea, the analysis continues to Step 2A, Prong Two
        - Step 2A, Prong Two: Integration into Practical Application

- If an AI invention is found to recite an abstract idea under Prong One, it can still qualify for patent protection if it integrates that abstract idea into a practical application under Prong Two.
    - The most effective way to demonstrate such integration is to show that the AI invention improves the functioning of a computer or another technology or technical field.
- Only if a claim fails both prongs of Step 2A does the analysis proceed to Step 2B

- Step 2B, examines whether the claim adds "significantly more" than the judicial exception itself

- Who is attempting to patent the invention?
    - The Patent Act requires that inventors must be natural persons; that is, human beings. *Thaler v Vidal*
        - Patent Act has defined an "inventor" as "the individual or, if a joint invention, the individuals collectively who invented or discovered the subject matter of the invention." *35 U.S.C. § 100(f)*
        - 'individual' ordinarily means a human being, a person." *Mohamad v. Palestinian Auth*
        - This begins and ends with the plain text. *Thaler v Vidal*
    - An AI (non-natural person) can also not be listed as JOINT Inventor. This is because an AI, following USPTO Guidance cannot CONTRIBUTE. Contribution is defined by the factors laid out in *Pannu v. Iolab Corp*
        - (1) contributed in some significant manner to the conception of the invention;
        - (2) made a contribution to the claimed invention that is not insignificant in quality, when that contribution is measured against the dimension of the full invention; and
        - (3) did more than merely explain the well-known concepts and/or the current state of the art
            - These factors must be met on every claim.
    - Interpretation of the *Pannu* factors from the USPTO
        - A natural person's use of an AI system in creating an AI-assisted invention does not negate the person's contributions as an inventor. YOU CAN USE AI
        - Merely recognizing a problem or having a general goal or research plan to pursue does not rise to the level of conception.
        - Reducing an invention to practice alone is not a significant contribution that rises to the level of inventorship.
        - A natural person who develops an essential building block from which the claimed invention is derived may be considered to have provided a significant contribution to the conception of the claimed invention even though the person was not present for or a participant in each activity that led to the conception of the claimed invention.
        - A person simply owning or overseeing an AI system that is used in the creation of an invention, without providing a significant contribution to the conception of the invention, does not make that person an inventor.
    - JAFFER notes that not permitting AI to file patents might not be a bad thing
        - If someone infringes on an AI's patent, how does the AI take action.
        - What if patent generating AI just became patent trolls
- Did the patent application disclose that an AI was used?

    - This doesn't matter for patenting per USPTO guidance.
    - This is different from the COPYRIGHT Office. In COPYRIGHT, the use of an AI must be made clear
    - This broadly suggests that the USPTO sees AI as a tool, nothing more

- Formal Guidance from the USPTO on AI Patents
    - Four Criteria (Same as above)
        - 1. The AI invention must be novel, meaning it cannot have been previously known.
        - 2. It should exhibit a sufficient level of inventiveness or non-obviousness that would not be apparent to an expert in the field.
        - 3. The AI invention needs to demonstrate practical utility and provide tangible benefits.
        - 4. Lastly, under Section 101 of U.S law, the AI innovation has to fall into recognized categories of patentable material.
    - Per federal laws on patents, abstract concepts along with mathematical calculations are beyond the scope of being patented. Similarly barred from obtaining patent rights are natural occurrences and fundamental scientific principles
    - Examples of Patentable AI Innovations
        - U.S. Patent No. 3,308,441 for an artificial neural network.
        - Developments such as data cleansing methods are potential candidates for patent protection.
        - USPTO Emphasizes - outlining the configuration of an AI model serves to set it apart from existing prior art by showing its uniqueness and inventiveness

Intellectual Property Production (Copyright)

- Is the copyright a valid copyright? Three elements
    - Originality: The work must be independently created by the author and possess a minimal degree of creativity (i.e., it cannot be copied).
    - Fixation: The work must take a tangible form (i.e., it can be perceived, reproduced, or otherwise communicated). Includes writings, recordings, a digital files, or any other physical medium.
    - Work of Authorship: Copyright does not protect ideas, procedures, concepts, titles, short phrases or other de minimis items. You cannot copyright a recipe, for example. You can copyright the spiritual journey that lead to that recipe on your cooking website
        - YOU DO NOT NEED TO REGISTER a copyright with the copyright office but there are increased protections for doing so.
- Is the copyright infringement claim valid?
    - For a copyright to be infringed, the copyright must be valid (see above)
    - Defendant must have infringed on the copyrighted work.
        - display of a copyrighted photograph, the broadcast of a copyrighted video, or the performance of a copyrighted play without authorization
        - Look for SUBSTANTAIL SIMILARITY
        - A person does not infringe a copyright merely by using information contained in a copyrighted work. *Feist Publications, Inc. v. Rural Tel. Serv. Co.*
    - Defendant must act willfully to violate the copyright. *MGM Studios Inc. v. Grokster*

        - This does not mean that defendant needs to be aware that they are violating a copyright. It just means that they need to have acted with purpose. *Sony Corp v Universal Studios*
    - Defendant acted to acquire Commercial Advantage or Private Financial Gain. *Sony Corp v Universal Studios*
        - It is not necessary to show that defendants actually made gains. *Sony.*
        - Posting copyrighted material on a public on a computer network, with knowledge that the copyright owner intended to distribute it, also meets this prong. *MGM Studios Inc. v. Grokster*
        - *Sony* and *Grokster* setup a tension. *Sony* is about Beta-max, which opens the door to facilitating unauthorized copying. But Sony prevails because there are other uses beyond unauthorized copying. *Grokster* facilitates unauthorized copying, but Grokster does not prevail because that is the only material use the service provides is unauthorized copying.
- Is the work in the public domain?
    - If the work is public domain there is no violation. *Klinger v Doyle Estates*
    - Copyright only exists for life of the author + 70 years.
- Is the reproduction subject to fair use protections? (Four Factors) – Cite *Warhol*
    - Check the analysis in *NY Times v Open AI* (motion to dismiss) and *Thompson Reuters v ROSS Technologies* (Summary Judgement)
    - Purpose and Character of the Use - Nonprofit, educational, and personal uses are more likely to survive scrutiny than commercial uses (but all can survive and fail)
        - Transformative uses, uses that result in the creation of a new work, with a new purpose and different character are favored as fair use over reproduction. The more transformative the less the other factors matter. *Warhol v Goldsmith.*
            - In Warhol a color change was insufficient to be transformative
    - The Nature of the Copyrighted Work - Factual works, published works and scientific articles are more likely to be considered available for fair use than are creative works. "Consumable" works, e.g. standardized tests, are not likely to be considered available for fair use. *Warhol v Goldsmith.*
    - Amount and Substantiality of the Portion Used – No bright line. The smaller the portion the better for fair use to succeed. Context matters along with purpose and character. Amount and substantiality is also a qualitative assessment. The key phrase is whether the portion used goes to the "heart of the work." *Wright v. Warner Books*
        - "[w]hat matters is not 'the amount and substantiality of the portion used in making a copy, but rather the amount and substantiality of what is thereby made accessible to a public for which it may serve as a competing substitute.'" *Thompson Reuters v ROSS*
    - Effect on the Potential Market or Value of the Work: Size of potential economic harm resulting from fair use. Again, evaluated in tandem with other factors
- Is the work parody or satire?
    - This is incredibly strong protection for fair use. *Warhol*
- Was AI use disclosed during the copyright application?
    - US Copyright Office requires the disclosure of AI use if it is more than *de minimis* use
    - This can void the copyright, but you can also refile
        - Note that this is different from patents. If a patent is rendered invalid, the game is up. Refiling copyright happens all the time.
- Was AI used to produce more than a *de minimis* amount of the copyrighted work?

- This is *Zarya of the Dawn* territory – the images from Midjourney are not copyrightable, but the text, characters, and other products of the artist are valid copyright. Autonomously generated content absent human involvement cannot be copyrighted
    - The issue is the randomness of the AI in production. It isn't like a camera, where the artist has complete control.
    - If there was more control, maybe this would be copyrightable. The issue is randomness, and because it isn't your own mental conception, it isn't copyrightable.
        - Textual prompts are suggestions
- Order and control matter tremendously. Photographs, for example, can be copyrighted because of the control the photographer has. *Burrow-Giles Lithographic Co. v. Sarony*
    - Postproduction on work in film that originally comes out of AI is copyrightable
    - This is based on the understanding of GenAI that is currently available.
    - CRITICAL: based on current technology
    - Look for SUBSTANTIAL HUMAN INVOLVEMENT

Infringement of Intellectual Property

- What is the theory of harm?
    - Direct infringement?
    - Contributory infringement?

- Does this the infringement constitute fair use?
    - See notes from above

- Do we have a *Sony* or *Grokster* situation?
    - Producers of technology who promote the ease of infringing on copyrights can be sued for inducing copyright infringement committed by their users. *MGM v Grokster*
    - The differentiation between *Sony* and *Grokster* is predominant use. Sony dealt with a claim based on distributing a product (Betamax VCRs) with alternative lawful and unlawful uses, with no specific intent to promote infringement. Grokster, on the other hand, involved the distribution of software specifically designed for peer-to-peer file sharing, where the predominant use was to infringe copyright.

- Is there a statute of limitations issue?
    - Because the statute of limitations is an affirmative defense, defendant bears the burden of establishing it. *RADesign*
    - The statute of limitations is 3 years, but the issue is when plaintiff became aware of the infringement through the exercise of reasonable diligence. *NY Times v Open AI*
    - There is no "a general duty to police the internet" to uncover infringement. *Parisienne v. Scripps Media, Inc*
    - Defendant must articulate why their behavior, if known to the plaintiff, should have put them on notice that there was infringement. *Cf. McGlynn.*
    - Sophisticated publisher argument – heightened expectation of plaintiff
        - 2nd Circuit has rejected this argument. *RADesign*
        - Could apply in the 9th Circuit

- Elements of a Direct Infringement Claim (Two Elements)
    - Make sure to reference the above on copyright protections
    - In *NY Times v Open AI* – Open AI does not succeed in motion to dismiss.
        - Microsoft does succeed in motion to dismiss

    - Copyright owner must demonstrate 1) that the defendant has actually copied the plaintiff's work; and 2) the copying is illegal because a substantial similarity exists between the defendant's work and the protectible elements of plaintiff's." *Yurman Design, Inc. v. PAJ, Inc*
    - Prong 1
        - This is a factual issue. Is there actually a copy out there somewhere?
            - This could be on a hard drive, but it more likely needs to be public
    - Prong 2
        - The standard test in determining substantial similarity is the 'ordinary observer test': whether an average lay observer would overlook any dissimilarities between the works and would conclude that one was copied from the other." *Nihon Keizai Shimbun, Inc. v. Comline Bus. Data, Inc*
            - This is not the exclusive provenance of a jury
            - Recall that facts cannot be copyrighted, but aggregations of facts can

- Elements of a Contributory Infringement Claim (3 elements)
    - In *NY Times v Open AI* – Open AI does not succeed in motion to dismiss. Microsoft does succeed in motion to dismiss
    - Copyright owner must demonstrate "(1) direct infringement by a third party, (2) that the defendant had 'knowledge of the infringing activity,' (3) and that the defendant 'materially contribute[d] to' the third party's infringement." *Dow Jones & Co., Inc. v. Juwai Ltd.*
    - Prong 1
        - See above in Direct Infringement notes above
    - Prong 2 (Scienter Requirement)
        - Second Circuit – Whether defendants objectively "know or have reason to know" of the direct infringement by third-party end users. *Gershwin Publ'g Corp*
            - More than a generalized knowledge of the possibility of infringement is required to meet the knowledge requirement. *Hartmann v. Apple, Inc*
            - Knowledge of specific infringement is not required. *Usenet.com, Inc*
                - These are the two ends of the spectrum. More than general, less than specific
        - Ninth Circuit – Liability for contributory copyright infringement requires that the defendant have possessed actual knowledge of or willful blindness to specific acts of infringement. *Ludvarts, LLC v. AT&T Mobility.*
    - Prong 3
        - To show defendant "materially contributed" to the infringement, the complaint must show that the defendant "encouraged or assisted others' infringement[] or provided machinery or goods that facilitated infringement." *Arista Recs. v. Lime Grp.*
            - In a digital space we are back to *Sony* and *Grokster*

Other Theories of Harm

- Is there a child involved?
    - The involvement of a child opens the door to COPPA. COPPA is the Children's Online Privacy Protection Act, which was designed to protect children under 13 from having their personal information collected online without parental consent.
        - Violations occur if websites or online services directed to children under 13 to fail to obtain verifiable parental consent before collecting, using, or disclosing their personal information.

- COPPA claims also open the door to an Unlawful Enrichment claim if the firm is unjustly profiting from the lack of granted consent. *AF v Character AI*

- Is there a product liability issue?
  - Example Case from class is *AF v Character AI*. Core issue in the case is targeting of children with a dangerous AI that induced anorexia and got a child to attack their parents.
    - Similar case: *Garcia v Character AI* – child killed themselves
  - Strict Liability Claim - offenses where a person can be held responsible for committing an act, even if they lacked INTENT or were NOT NEGLIGENT. A manufacturer is liable under a strict liability product defect claim if the product was in an unreasonably dangerous defective condition when put to a reasonably anticipated use, and the plaintiff was damaged as a direct result of such defective condition as existed when the product was sold. The elements for strict liability defective product are identical to strict liability defective design.
    - Necessary Elements: *Dorgan v. Ethicon*
    - 1) the defendant sold the product in the course of its business;
    - 2) the product was then in a defective condition, unreasonably dangerous when put to a reasonably anticipated use;
    - 3) the product was used in a manner reasonably anticipated;
    - 4) the user was damaged as a direct result of the product.
  - Common Law Negligence: legal doctrine where individuals have a duty to exercise reasonable care and skill to avoid causing harm to others. To prevail on a claim for negligence under common law, a plaintiff must establish four elements: *Amick v. BM & KM*
    - Duty of Care: Legal obligation a to exercise reasonable care in their actions or omissions to avoid causing harm to others.
      - This can come from i) established precedent, ii) PROXIMITY between the defendant and the plaintiff, or iii) FORESEEABILITY (reasonably foreseen action)
    - Breach of Duty: Occurs when a party fails to meet the required standard of care, either through their actions (commissions) or their failure to act (omissions).
      - defendant's actions or inactions fell below the standard of care expected of a REASONABLE PERSON in SIMILAR CIRCUMSTANCES
    - Causation: The breach of duty must be the direct cause of the harm or damage suffered by the injured party. Can be either
      - "But for" causation (if the breach hadn't occurred, the harm wouldn't have happened)
      - Proximate causation (the breach was a foreseeable cause of the harm).
    - Damages: The injured party must have suffered actual harm or loss as a result of the negligence (e.g. physical injuries, property damage, economic losses)
  - Aiding and abetting liability – unlike negligence claims, aiding and abetting liability does not require the existence of, nor does it create, a pre-existing duty of care to a third party nonclient. Rather aiding and abetting liability is based on proof of scienter - the defendants must know that the conduct they are aiding and abetting is a tort. *Restatement (Second) of Torts § 876*. Two elements to make this claim. Plaintiff must show: *Newby v. Enron Corp.*
    - (1) that the aider and abettor knows that the party he is aiding is breaching a duty it owes to another and
    - (2) that the aider and abettor gave substantial assistance to the primary tortfeasor.

- In *Character AI*, aiding and abetting claims were brought against Google because of the support given to Character AI in development and the fact that they previously had shut down the project due to the dangers of it

- Is there a discrimination issue (e.g. bias against a protected class)?
  - Is the AI an AGENT of the firm?
    - Agency Liability - legal responsibility placed upon a principal (e.g., employer) for the actions, contracts, and torts (wrongful acts) of their agent (e.g., employee) when the agent is acting within the scope of their authority.
    - In *Mobley v Workday* the court determines that Workday is an AGENT of the firm because it takes over "functions [that] are traditionally exercised by an employer." *Williams v. City of Montgomery*
      - Distinction is critical. Without Agency Liability the firm would only be liable for the discrimination the AI / AI Vendor informs them of
  - Is the AI / AI Vendor an Employment Agency?
    - An Employment Agency is one that is curating jobs and matching them to people.
    - If the AI was an Employment Agency, only the AI Vendor would be liable
    - In *Mobley,* Workday (AI) was not curating and matching! Workday is a screening mechanism.
  - Is there Disparate Impact? Plaintiff must show: *Bolden-Hardge v. Off. of Cal. State Controller*
    - (1) a significant disparate impact on a protected class or group;
      - Passes in *Mobley:* Workday is distinguishing between who you should interview and who you shouldn't. Judge rationalizes that you can infer from the fact that he was rejected by 100 times and there were jobs that he was clearly qualified for
      - OBSERVATION: I find this peculiar because there is no counterfactual comparison group. Instead it was based on the number and speed of rejections. A counterfactual would be stronger evidence
    - (2) identify the specific employment practices or selection criteria at issue; and
      - Workday, which is an agent, meets this prong in *Mobley*
    - (3) show a causal relationship between the challenged practices or criteria and the disparate impact.
      - Very quick reactions over 100 of times, and Workday is making agenic decisions about WHO is interviewed or not. Rejections are happening so early in the morning. The AI is CAUSING THIS
      - Consider "but for" or proximate causation as needed
  - Has there been intentional discrimination? Four prongs *Sheets v. City of Winslow*
    - (1) Plaintiff must be a member of a protected class;
    - (2) Plaintiff must be qualified for the position(s);
    - (3) Plaintiff must have experienced an adverse employment action; and
    - (4) similarly situated individuals outside the protected class must be treated more favorably than Plaintiff, or other circumstances surrounding the adverse employment action give rise to an inference of discrimination.
      - In *Mobley,* plaintiff succeeds on the first three prongs. Fails on the final.
      - Stating a claim for disparate treatment requires pleading facts giving rise to an inference that the employer INTENDED to discriminate against the protected group. *Liu v Uber Techs. Wood v City of San Diego.*

    - Were reasons for the unfavorable outcome given that could justify the decision?
        - See *Louis v SafeRent*. Liability could probably have been avoided.
        - This should be specific (e.g. career length, criminal record, training)
    - Was there an appeals process where deficiencies in algorithmic assessment could be addressed?
        - See *Louis v SafeRent*. Liability could probably have been avoided.
        - The appeals process would theoretically permit plaintiff to correct information errors or information gaps (the voucher in *SafeRent*)

- Is there a defamation issue?
    - Put yourself in mind of *Walter's v Open AI*. Hallucination claiming that Walters was accused of embezzlement (claims of active litigation) which was completely false.
    - Four elements of a defamation claim per *Paterson v. Little, Brown & Co. Mark v. Seattle Times*. Plaintiff bears burden of establishing a *prima facie* case on all four elements.
        - 1) falsity – statement must be demonstrably false. TRUTH is a complete defense
        - 2) an unprivileged communication – this is publication to a third party. The statement needs to be communicated to someone else.
        - 3) fault – hinges on public or private persons
            - Public Figures – "actual malice" is needed. *NY Times v Sullivan.*
                - A public person is one who has achieved above average (pervasive) fame or notoriety (e.g. celebrity, political figure, athlete)
            - Private Persons – negligence standard. Failure to exercise reasonable care
                - See above
        - 4) damages – actual or presumed. Harmed reputation, economic losses. These are frequently presumed based on other elements.
            - Actual damages – look for a particularized economic or reputational loss
            - Presumed – law has assumed a harm to reputation or compensation without specific proof, e.g. defamation *per se*
    - Keep in mind FERRARO suggestion for defective liability here. See elements of Strict Liability claim above (Duty, Breach, Causation, Damages)
        - Intuition – An AI that "defames" someone mike open the door to defective liability

- Is there a facial recognition issue?
    - Not entirely sure how to prep this. Refer to *Woodruff v Oliver* below. Check the FTC Enforcement Action against Rite Aid. If we see this, it will probably fall under racial discrimination or discrimination against some other class
    - How to avoid such issues / minimize liability?
        - Human in the loop on decision making
        - Downgrade odds with using older or low quality photos
        - Cross reference with other public information
        - FERRARO brings up lighting structure of images. Others brings up "photo focal length"
        - ACLU argues that neither a FRT (Facial Recognition Technology) result, nor an eyewitness identification from a photo array based on an FRT result, should supply probable cause for an arrest. FRT results are fundamentally unreliable and display higher rates of false matches for people of color, women, and young adults
    - For corporations using FRT, see *In the Matter of Rite-Aid Corp.* FTC argues this is a deceptive trade practice (using FRT) The crux is largely the same.

        - Inform consumers
        - Properly train employees
        - Investigate complaints and false positives
        - Enable third party oversight
        - Delete older biometrics
    - Violation of the FTC Act - unfair or deceptive acts or practices in or affecting commerce are unlawful. FTC must show
        - 1) a representation, omission or practice that is likely to mislead (misrepresentation or omission);
        - 2) that consumers were misled while acting reasonably under the circumstances;
        - 3) lack of reasonable consumer ability to avoid the injury; and
        - 4) substantial injury
        - Differentiating Unfair and Deceptive Acts
            - Unfairness: unfair if it is not justified by any offsetting benefits to consumers or competition. (RITE-AID!)
            - Deception: if an act is likely to mislead consumers in a material respect.

Intimate Harms

- General Background and Policy Issues
    - Resources if anyone needs them - https://www.dhs.gov/know2protect
    - At present, 96% of deep fakes on the internet are pornographic. And the vast majority of those are non-consensual. The term deepfake = deep learning + fake
        - Deepfake comes from a subreddit
    - This includes child pornography. Term of art is Child Sexual Abuse Material (CSAM)
        - In 2014, there were 1mm reports
        - Up to 23mm reports annually in 2022
            - These are challenging not only due to the production of illicit material but frequently DHS attempts to find these children and they do not exist (completely fictionalized)
    - The DOJ has noted production of deepfakes of minors engaging in sexual acts using Stable Diffusion (Open AI Product)
        - Wisconsin Indictment of Anderegg (*US v Anderegg*)
            - Production of more than 13,000 images using Stable Diffusion
            - Distribution of those images to minors
            - Instruction of how to create those images to others (including minors)
    - The US does not have a general comprehensive federal privacy act. No equivalent to GDPR
        - All 50 states have their own and they are inconsistent
    - Things you could address
        - Notice and disclosure – requiring consent to collect information
        - Opt out requirements – users can opt out of business requirements
        - Deletion and minimization requirements – right to be forgotten
    - This could be implemented by the FTC under unfair or deceptive trade practices

- Is there an issue with people creating fake NCII?
    - Relevant Statutes (*People of California v Sol ECOM et al*)
        - California Civil Code (CCC) 1708.86(b)(1) prohibiting the creation and intentional disclosure of nonconsensual sexually explicit images, or aided and abetted violations

        - CCC 1708.85(a) prohibiting the intentional distribution of nonconsensual depictions of intimate body parts, or aided and abetted violations
        - CCC 647(j)(4) prohibiting the intentional distribution of nonconsensual depictions of intimate body parts of an identifiable person, or aided and abetted violations
        - FEDERAL - 15 U.S.C. § 6851(b)(1) prohibiting the knowing or reckless disclosure in interstate commerce of intimate visual depictions of identifiable persons, or aided and abetted
            - This is the Violence Against Women Reauthorization Act of 2022. Civil action in federal court against someone who shared intimate images, explicit pictures, recorded videos, or other depictions of you without consent
            - Includes sharing those intimate images through technology, such as the internet or social media.
        - Unfair business acts and practices (CCC 17200). Creating nudified images constitute unfair business practices because they offend established public policy, the harm they cause to consumers greatly outweighs any benefits associated with those practices, and they are immoral, unethical, oppressive, unscrupulous and/or substantially injurious to consumers.
            - There are additional violations for such conduct for websites that create child images
    - Legal test for consumer protection violations (California). *Sepanossian v. National Ready Mixed Concrete Co.*
        - Plaintiff must demonstrate that members of the public are likely to be deceived. The notion is that consumers who witness the NCII may think this is real
        - TEST: reasonable consumer standard
            - A 'reasonable consumer' is 'the ordinary consumer acting reasonably under the circumstances', and 'is not versed in the art of inspecting and judging a produce in the process or its preparation or manufacture... .' *Colgan v. Leatherman Tool Group, Inc.. Cochrane v. Elliott*
            - This is an objective standard: The reasonable consumer standard is not about the specific consumer who might be affected, but about a hypothetical "reasonable consumer".
            - The standard focuses on whether the advertisement or labeling has a "likelihood of deception" or "capacity to mislead" a reasonable consumer.
            - The standard is not about proving that a specific consumer was actually misled, but rather about whether a reasonable consumer would be misled.
        - See additional legislation below
            - Federal - TAKE IT DOWN
            - Hawaii Revenge Porn and NCII statute
            - California deep fake statutes

National Security

- What are the core issues we are dealing with?
    - CBRN – Chemical, Biological, Radiological, and Nuclear Weapons
    - Defenders dilemma – assumes that defenders are waiting for attacks to happen and then respond. This reactive stance allows attackers to define the terms of engagement and puts the defenders in the position of always playing catchup
        - Attackers just need to win once. Defenders need to win every time

    - Cybersecurity – encapsulates both offensive and defensive concerns
    - Disinformation – false information which is meant to mislead parties
    - See above list in Definitions and General Background

- History of AI CBRN issues
    - EO 14110 was issued under Biden and rescinded by Trump. I think where FERRARO wants us to go is that there is a balancing act in the principles. We want to accentuate the good and ameliorate the bad. The principles are out there to provide a framework to start down this path
    - The EO was written following the early AI frameworks from both Biden and Trump I. Broad guidance sort of stuff. Ben Buchanan is the author.
        - There is also the National Defense Authorization Act which had some AI stuff in it.
        - NDAA establishes a national network for
            - Microelectronics research and development,
            - Identification of objectives and establishment of performance metrics (how AI can be used and where we need to be upskilled)
            - Establishment of executive level civilian and military education, briefings, AI training, and education at the US Naval Community College
            - Reporting requirements to the National Security Commission
            - Assessments of cyber-posture, and comparative readiness with adversaries
        - This EO draws heavily from the Blueprint for an AI Bill of rights
        - After ChatGPT emerged, the new goal was to take a broader governance approach from the Executive Branch.
    - There is one exception to the EO approach. This was a requirement that AI developers (under the Defense Production Act) disclose to the Feds that they are working on such things
        - Everything else is voluntary by and large
    - EO Principles
        - AI must be safe and secure – unintended consequences. Misuse. Misappropriate by malicious actors. Pre-deployment testing
        - Responsible innovation, competition, and collaboration – Maintain US leadership in this space. Investment in education, training, and novel intellectual property issues, antitrust
        - Supporting the American Worker – training and education, collective bargaining,
        - Equity and Civil Rights – obvious one here. In line with Biden political priors. "Holding AI developers" to standard. This is done by language to BUILD TRUST
        - Transparency and Consumer Protection – model understandability and precluding harms to consumers
        - Privacy, civil rights, and civil liberties – no bias in decision making
        - Accountability in the use of AI by the federal government – managing and building capacity / infrastructure to use this properly
        - The US should lead the way, but also partner with international partners and the private sector. US Leadership
    - MY CONCERN, this is piffle. FERRARO pushes back. There is a presumption that AI is economically efficient and valuable but there is no benchmark for that claim. Takes the efficiency for granted

- Dual Use Foundation Models

    - The term "dual-use foundation model" means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:
        - (i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;
        - (ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or
        - (iii) permitting the evasion of human control or overnight through means of deception or obfuscation.
    - Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities.
    - Dual use models are critical. Differentiate the TECHNOLOGY (thing) and USE (implementation). There is the thing, and there is the use. In a lot of regulatory circles there is a focus on the THING. In other places there is the USE. Use is what you are trying to do with it. THING is technical factors (speed etc.)

- Biological Uses
    - The concern with biologics and AI is that there are a lot of known unknowns
    - There are use cases in chemical and biological settings for AI (e.g. drug development), and we might presume that it might get applied in warfare. Via:
        - Democratized use of open access tools
        - Munificence of open source data
        - AI's ability to process data very quickly
    - Intuition is that (via dual use) the AI could turn a good thing (making drugs) into a bad thing (accelerating a virus)
    - Mild mental gymnastics to move this towards chemical, radiological, or nuclear

- Governance issues and challenges
    - Traditionally, the Cybersecurity and Infrastructure Security Agency (CISA) would oversee Chemical manufacturing through facility inspections (via Chemical Facility Anti-Terrorism Standards (CFATS)). This has expired.
    - Compounding issue: There is presently a lack of established upstream and downstream regulatory frameworks. No uniform framework (Swiss Cheese issue)
        - Upstream – inputs, data, training, research, computing power
        - Downstream - Sensitive data, export controls, intellectual property
    - Compounding issue: How do we regulate in a world of open source? Open source can be used by any actor, benevolent or malicious

- Cybersecurity considerations
    - Defenders dilemma
        - There are lots of good examples of AI use in cyber security (positive and negative). The issue is that on defense you need to bat 1000. An attacker only needs to penetrate once. The core argument for AI in context is that AI will improve attacker and defender abilities, but defenders gain more by automating defense.
        - How so?

- Reasoning – AI can scan millions of lines of code quickly to identify holes in code (be it established code or malware)
- Learning – AI can get better over time, without being coded to do so
  - Example, Google has an anti-money laundering AI tool which limited false positives and increased true positives
- Speed – AI moves at machine speed rather than human speed
- Scale – AI operates at scale, humans don't
  - Google Safe browsing – identifying bad websites at scale quickly

- Potential Policy Suggestions (that are kindly in line with Google's business interests)
  - Secure by design (SBD) - Secure AI from the ground up. This should not be an add on. It should be a design principle
    - Zero trust – never trust, always verify
    - Implementing passwords and multifactor authentication
    - Security is integrated, not overlaid ex post onto a system
  - Preserve the ability to train on public data and augment with private data
    - Brad is skeptical of the argument that Open AI wouldn't be getting sued by the NY Times if they made a better Norton Antivirus
  - AI Training for users
  - Promote AI security for critical infrastructure rather than prohibit it
    - The Good – AI is a solid defender against attackers at scale
    - The Bad – what if the AI gets corrupted
  - Pursue research in i) design and build and ii) system safety in use
- The core of the defender's dilemma is that there are marginal improvements of capabilities for both sides. But the margin for defenders is larger. Benefits:
  - Discovery and getting the information into MITRE's known vulnerability database
  - Improved the time to recovery, i.e., reduced the time to exploit vulnerability
  - Improves time to proliferate patches that solve exploited vulnerabilities

- Regulatory guidance the Cybersecurity and Infrastructure Security Agency (CISA) offered on how to deploying AI systems securely. CISA is being dismantled by the current administration. A secure system will have the following:
  - Secure the deployment environment: coded in a secure location
  - Ensure the robust deployment environment infrastructure: hardware and software that the AI is deployed on is secure
  - Harden deployment environmental configurations: physical security of location
  - Protect distribution networks from threats: penetration can come from physical or remote attackers
  - Validate system use before and during use: red teaming etc. Validation of performance and security will be ongoing
  - Secure exposed APIs – don't leave the back door open for a stack overflow issue
  - Actively monitor model behavior
  - Protect model weights
  - Secure operation and maintenance
  - Enforce strict access control
  - Ensure user awareness and training
  - Conduct audits and penetration testing

        - Implement robust logging
        - Update and patch regularly
        - Prepare rollbacks – Grandfather, father, son
        - Plan secure delete capability – pull the plug

- Disinformation Considerations
    - *In the Matter of Steve Kramer*
        - FCC action on robocalling. In the great state of New Hampshire, there was a spoofed voice of Biden calling voters and telling them that they didn't need to vote
        - Action is brought by the FCC, the State AG, and other parties
        - Penalties:
            - There is a 1mm fine, an injunction, and a monitoring program
        - Causes of action - FCC
            - FCC isn't worked up about the deep fake, it's about AI spoofed VOIP calls from hidden numbers. This is a violation of Know Your Customer and Know Your Upstream Provider Rules that permit telecom firms to vet traffic. Lingo Telecom (Kramer's Firm) circumvented this
        - Causes of action - League of Women Voters
            - Section 11(b) of the Voting Rights Act
            - Intimidation, threats, coercion
            - "There is limited precedent discussing what "threaten" and "intimidate" mean in the context of Section 11(b). But the plain meaning of those terms, the few cases that interpret Section 11(b), and cases that interpret the same or similar language in other civil rights statutes all indicate that intimidation includes messages that a reasonable recipient, familiar with the context of the message, would interpret as a threat of injury - whether [*18] physical or nonviolent - intended to deter individuals from exercising their voting rights. Threats, intimidation or coercion may take on many forms. As the Court explained previously, actions or communications that inspire fear of economic harm, legal repercussions, privacy violations, and even surveillance can constitute unlawful threats or intimidation under the statute." *Nat'l Coal. on Black Civic Participation v. Wohl*
        - Causes of Action - State charges (AG) of felony voter suppression and misdemeanor impersonation of a candidate
    - The issue with deep fakes: people can identify deep fakes, that's not the problem. They problem is that these fakes are part of a broader gestalt of misinformation
        - Received wisdom is that deep fakes are expensive to make, and are no more effective as old-fashioned disinformation campaigns
        - Reinforcing existing narratives is the key issue if you want to run a disinformation campaign
        - There's also a liar's dividend issue: when people learn that deep fakes are increasingly pervasive and realistic, false claims that real content is AI-generated become more persuasive too.
    - Big Picture on disinformation
        - Do you want to limit creation, do you want to limit exposure, do you want to limit belief in fake or real information, or do you want to limit action on that faked media?

        - What problem are you trying to solve? Because the problem you are trying to solve effects the policy you set

AI In Warfare

- What are the potential uses of AI in warfare? Congressional Research Report.
    - KEEP IN MIND THAT THE FOLLOWING WILL BE BOTH ADVESARIAL AND DEFENSIVE. WHAT WE CAN DO, THEY CAN DO.
    - Intelligence Surveillance and Recon – AI can process more data (video) and assess with some level of accuracy what is going on (if it is properly trained). Uses:
        - It is unclear if humans or AIs have a better idea of spatial relationships and their intrinsic relationship. X is happening in Riyadh, what is happening in Doha. This spatial reasoning based on intuition is challenging
        - Friend vs Foe identification
        - There are core issues with ground truth in surveillance. Training models is hard, but human intuition is also flawed
    - Logistics and Maintenance – including sensors in aircraft. Predicting when parts will fail and how to logistically get the parts to the individual pieces
        - This could also work on partial heuristics as well. A partial heuristic is a decision-making shortcut or mental rule that only provides a piece of the information needed to make a complete decision.
            - There are tertiary benefits, other parts that are likely to fail soon could also be repaired when the asset is undergoing maintenance (in addition to the primary part). Think of the US Steel example.
        - Modeling scenarios for warfare. Combat readiness overall, but these plans are usually broad. The AI might help develop plans that are more specific and account for idiosyncrasies associated with combat readiness (e.g. good weather / bad weather) can yield different course of action. More thorough modeling
        - Efficient loss or efficient failure – identifying which aircraft should be repaired. Identifying which assets should be cannibalized. This is a resource allocation game with limited resources
    - Cyberspace Operations – advancing defensive and offensive cyberspace operations
        - Defenders Dilemma – Detecting and identifying vulnerabilities. See above
        - Improved offensive operations. Attacking mission critical adversarial resources
        - Identifying network behavior and packet movement (larger scan approach)
        - Learning new threats that are incoming (net new capabilities)
        - All of these are both offensive and defensive, the AIs are playing chess against AIs
    - Information Operations and Deep fakes
        - Detection and creation of misinformation - generate false news reports, influence public discourse, erode public trust, and attempt to convert diplomats into assets
        - Creating information profiles of exploitable human assets - servicemembers, suspected intelligence officers, government officials, or private citizens
            - Think about PRC Hacks – TikTok, OPM hack (SF-86 data), and Experian hack. This is a social engineering and exploitation game
        - Both offensive (social engineering) and defensive (identification of weak points within national assets) options
        - Human intelligence and counterintelligence (Potential)
    - Command and Control - centralize planning and execution of air-, space-, cyberspace-, sea-, and land-based operations

        - Single platform for coordination across services.
        - KEEP IN MIND: There are obvious human resistance issues (people don't always want to share), perceived usefulness, perceived ease of use, political frictions
        - KEEP IN MIND: Compartmentalized Information issues. The theory is that this could improve through the centralization of information.
            - Post 9/11 pivoted from need to know to sharing information. Post Snowden and Manning people are getting skittish. Ongoing tension
            - DC Yo-yo game of overcorrection
        - KEEP IN MIND: Tension across military and intelligence here – who is running the show. Is it DNI? Is it DOD? Who wants to share
            - JAFFER talks about the acceleration of command and control operations in Iraq which stems from JADC2 (Joint All Domain Command and Control)
            - The notion is that we can improve military operations through information sharing across branches seamlessly, but also improve operations with real time field information incorporation (on the ground data)
    - Semiautonomous and Autonomous Vehicles – potential for improved decision making, limited loss of life, lack of risk to human operators

- Ethics of Lethal Autonomous Weapons Systems (LAWS)
    - DODD 3000.09 defines LAWS as weapon systems that, once activated, can select and engage targets without further intervention by a human operator
        - Note – there is no agreed upon definition but this one is useful
    - There are some core ethical issues here (Langley Article).
        - Discrimination Principle – they cannot ensure that the system will not harm civilians, and if they fail, who do you hold accountable. Flat principle failure
        - JAFFER pushes back – if we think about this as a net welfare issue, and think about the counterfactual world, we don't know if discrimination will improve or not.
    - Distinctions of Systems
        - Human On the Loop – autonomous weapon system that is designed to provide human operators with the ability to intervene and terminate engagements before unacceptable levels of damage occur
        - Human in the Loop – systems intended to only engage individual targets or specific target groups that have been selected by a human operator. AFFIRMATIVE decision to engage / destroy made by the human
        - Autonomous Systems - systems, once activated, that select and engage targets without further intervention by a human operator. Weapons free engagement
    - Types of LAWS Systems (Examples)
        - Autonomous Sentry Systems – Koreans using automatic gunfire on the DMZ. Phalanx System. AEGIS systems on boats looking for missiles, drones, and planes.
            - Human on the Loop – person can turn on and turn off
            - Concern – this is moving from ships to in field deployment. Moving from defensive to offensive capabilities at scale
        - Autonomous Killer Robots – slaughter bots – clone army from the separatists
        - Autonomous Drones and Swarms – distributed brain drones (hundreds) that can be deployed on a battlefield
    - Core Ethical Issues

        - These last two (Killbots and Swarms) are Langley's concern. Capabilities being built at scale by the US and adversaries. Once you let them run, they cannot discriminate between friend and foe. NOTE: Inability to discriminate efficiently is speculative.
            - Entirely possible, but it is also possible that the killbot will get better over time. Potentially even better that a human with practice.
        - Regardless, if you leave the system to operate autonomously, you open yourself to some ethical problems. Rules of war violation
            - The ethicist based on Just War Theory would call this a *jus in bello* issue. "Justice in war" or "right conduct in war," is a set of principles and rules that dictate the conduct of armed conflict. It focuses on how a war should be fought, regardless of whether the war itself is just or not.
            - The objectivist ethical reasoning (Bentham) would say that if the killbot kills the same number of people as a human would, we are at a welfare wash
                - Jeremy Bentham would say the same thing
    - *Jus in bello* (Justice in War): Refers to the rules that govern the conduct of war, aiming to minimize harm and suffering. Key Considerations
        - Discrimination: Combatants must be distinguished from civilians, and attacks must be directed only at legitimate military targets.
        - Proportionality: The harm caused by military actions must be proportionate to the military advantage gained.
        - Necessity: Only the force necessary to achieve the military objective should be used.
        - Responsibility: Those who violate the rules of war should be held accountable.
    - *Jus ad bellum* (Justice of War): Refers to the conditions that must be met before a war can be considered just. Key Considerations
        - Just Cause: The war must be fought for a just reason, such as self-defense, or to prevent or punish grave injustice.
        - Legitimate Authority: The decision to go to war must be made by a legitimate authority, such as a duly constituted government.
        - Right Intention: The war must be fought with the right intention, aiming to achieve a just outcome, not for selfish gain or conquest.
        - Last Resort: War should only be considered as a last resort, after all other peaceful means of resolving the conflict have been exhausted.
        - Reasonable Chance of Success: There must be a reasonable chance of achieving the war's objectives.
        - Proportionality: The use of force must be proportionate to the cause of the war and the intended outcome.

- Ethical Conundrum: What about killbots (PRC, Russia) versus killbots (US)?
    - Base supposition of this discussion is that killbots v killbots is an inevitability.
    - GREENWOOD SPECULATION: Killbots v killbots is just a property issue.

- Current DOD policy on LAWS (lethal autonomous) – D3000.9 requires
    - Control - all systems, including LAWS, must be designed to "allow commanders and operators to exercise appropriate levels of human judgment over the use of force."
        - Appropriate is not a one size fits all. It is flexible based on the system
        - Control is also a flexible term
    - Training - adequate training [tactics, techniques, and procedures] and doctrine are available, periodically reviewed, and used by system operators and commanders to understand the

functioning, capabilities, and limitations of the system's autonomy in realistic operational conditions
    - Review Process - requires that the software and hardware of covered semi-autonomous and autonomous weapon systems be tested and evaluated to ensure they:
        - Function as anticipated in realistic operational environments
        - Be sufficiently robust to minimize the probability and consequences of failures
    - Senior-level review - In addition to the standard weapons review process, a secondary senior-level review is required for covered autonomous and semi-autonomous systems.
        - This is done under the purview of the Under Secretary of Defense for Policy (USD[P]), the vice chairman of the Joint Chiefs of Staff (VCJCS), and the Under Secretary of Defense for Research and Engineering (USD[R&E])
        - This can be waived for "urgent military need"
        - There is a working group to inform senior level review
    - Congressional notification - the Secretary of Defense is to notify the defense committees of any changes to DODD 3000.09 within 30 days
    - THERE IS NO SCENARIO where humans are OUT of the loop. Limited to ON the loop
        - Potential flaw: if an adversary knows this is our policy, and humans in / on the loop are required, they gain a military advantage
            - Mike Tyson everyone has a plan until they get punched in the face
        - Huge ethical back and forth in class.

- USS Vincennes Iranian shooting incident
    - Not sure where else to put this. Just keep it in mind
    - The USS Vincennes incident refers to the 1988 shootdown of Iran Air Flight 655 by the USS Vincennes. The incident occurred during the Iran-Iraq War when the Vincennes, mistaking the civilian Airbus A300 for an Iranian F-14 fighter, fired two missiles, killing all 290 people on board.
    - When the Vincennes picked up the plane on radar, it was automatically identified as an aircraft 'assumed hostile' and tracked by the Aegis computer system on the cruiser

- Arguments in favor of the use of LAWS (Etzioni and Etzioni)
    - Military advantages
        - LAWS act as a force multiplier. Fewer warfighters are needed for a given mission, and the efficacy of each warfighter is greater.
        - LAWS expand the battlefield, allowing combat to reach into areas that were previously inaccessible.
        - LAWS can reduce casualties by removing human warfighters from dangerous missions.
        - LAWS are better suited than humans for 'dull, dirty, or dangerous' missions
        - LAWS can operate faster than a human warfighter
        - LAWS are not subject to physical limitations - high-G maneuvers and the intense mental concentration and situational awareness required of fighter pilots make them prone to fatigue and exhaustion
        - Programmed randomness (limited predictability)
        - Superior perception (limitations of the human's ability to sense)
        - Planning and learning
        - Multiagent coordination – central planning
    - Fiscal Advantages

        - "Each soldier in Afghanistan costs the Pentagon roughly $850,000 per year." Conversely, the TALON robot—a small rover that can be outfitted with weapons, costs $230,000
    - Moral Justifications
        - No self-preservation instinct, potentially eliminating the need for a "shoot-first, ask questions later" attitude.
        - No clouding of judgement by fear or hysteria
        - Robots could be more relied upon to report ethical infractions they observed than would a team of humans who might close ranks
        - Elimination of battle fatigue / PTSD / mental trauma

- Arguments against the use of LAWS (Etzioni and Etzioni)
    - Opposition on moral grounds
        - Currently AI combat is infeasible but this is a Pandora's Box. It will set off another nuclear arms race
            - Both offensive and defensive issues here, but no one questions the ethicality of a missile defense shield. The issue is purely framed offensively
            - I do wonder about a Dr Manhattan situation (Watchmen)
        - Lack of current international framework for use, no one agrees on what is permissible and what is not permissible
        - Potential for high levels of collateral damage
        - Delegation of life or death decisions to nonhuman agents (Discrimination)
        - Accountability (also distinction, *jus in bello*)
        - Articulation of objective functions (fake AI override)

- The Fake AI Override
    - Example of mild hysteria on these things
    - USAF officer talks about a simulation where the drone is supposed to target and kill certain topics. Drone got points for killing targets. Operator tells it not to kill a target. Drone waffles, then kills the operator, and then kills all the targets
        - Completely bullshit story – thought experiment was claimed. No one knows

- Other things to keep in mind
    - 1) Sensors - so much of all of what is going on is based on sensors and the accurate collection of information. Hinges on accurate data collection
    - 2) Tension between targeting and technology – building the machine v how you use it
    - 3) This (AI and warfare) is all being developed in real time in Ukraine right now
        - AI on the edge is happening right now
        - You also have a country that purportedly adheres to rules of war (Ukraine) v one that purportedly doesn't (Russia). There is contention here
    - The Etzioni article assumes that you will program a rational machine. You might not do this

State Level AI Regulatory Efforts

- FERRARO recommendation. State regulation is where the action primarily is. As you go through the statutes, think about them thematically.
    - How is the object of the regulation defined?
        - Is it the thing (what defines AI, 1.21 gigawatts)?
        - Are they concerned with how it is used?
    - What are the exemptions to the regulation?

    - What is required of the parties?
    - Whom is targeted (within the ambit)?
    - Who has enforcement authority (agency, AG, private parties)?
    - Under what authority / what is the justification?
    - By when (timing horizon)?
    - What are the remedies / penalties for non-compliance?
    - What are the exemptions?
    - What are targets of the regulation required to do?

- California Statutes
    - There's some intense disagreement about whether this will yield positives or negatives. There was a slew of these (almost 40). Article highlights 18 or particular import
    - EFFECTIVE DATE: 2026
    - California requires you to disclose which datasets you are using, how many data points you use, how you use them, where they are sourced from, if they are copyrighted, when the model was trained.
    - DEFINITION OF AI: an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments
    - AMBIT: California's AI laws primarily affect AI developers, businesses using AI systems, and state agencies utilizing generative AI. These include both those that develop or substantially modify existing AI systems
        - SUBSTANTIAL MODIFICATION means a new release or version that materially changes performance. RETRAINING OR FINE TUNING. This is wide encompassing. This is a big ass set of laws.
        - EMPHASIS: Fine tuning is substantial modification…
    - ENFORCEMENT: There are no private rights of action here. Everything is enforced by the AG or regulatory bodies.
    - Individual Laws:
        - AI Risk Management: SB-896, which mandates that California's Office of Emergency Services (CalOES) conduct risk analyses regarding generative AI's potential dangers. The law requires collaboration with frontier AI companies like OpenAI and Anthropic
        - AI Healthcare: AB-3030 requires health care providers to disclose when they use generative AI to communicate with patients, particularly when the messages contain clinical information
        - AI Healthcare: SB-1120 establishes limits on how health care providers and insurers can automate their services, ensuring that licensed physicians oversee the use of AI tools in these environments.
            - So, if you are going to automate SERVICES they must be done under the auspices of a physician.
            - Denials or delays must be made by a physician
            - Open to inspection and compliance
        - Privacy and AI: AB-1008 extends the state's existing privacy laws to cover generative AI systems.
            - If an AI system exposes personal information—such as names, addresses, or biometric data—businesses will be subject to restrictions on how they can use and profit from that data.

        - Any PII will be subject to restrictions on use on how the firm can profit.
        - AI systems must adhere to the same standard as everyone else in using PII
    - Watermarking AI Generated Content: SB-942 requires generative AI systems to disclose that the content they create is AI-generated.
        - This will be done through "provenance data" embedded in the content's metadata (e.g. OpenAI's DALL-E must tag in their metadata indicating they were generated by AI).
        - The watermark must be detectable
        - This requires the COMPANY to put the watermarked meta-data on an image, not necessarily a private person. The COMPANY is the regulated entity
    - AI and Education: AB-2876 requires the California State Board of Education to consider AI literacy when developing curriculum frameworks for subjects like math, science, and history
    - AI and Education: SB-1288 requires California superintendents to form working groups to explore how AI is being used in public education and identify potential opportunities and challenges.
    - Robocalls: AB-2905 requires robocalls to disclose when they use AI-generated voices
        - Goal is to prevent the NH situation (*In the Matter of Steven Kramer*)
    - Deepfake Pornography: AB-1831 expands existing child pornography laws to include content generated by AI systems.
    - Deepfake Pornography: SB-926 makes it illegal to blackmail individuals using AI-generated nude images that resemble them
    - Deepfake Pornography: SB-981 requires social media platforms to establish reporting mechanisms for users to flag deepfake nudes.
        - Platforms must temporarily block such content while it is under investigation and remove it permanently if confirmed as a deepfake.
    - Election Deepfakes: AB-2655 mandates that large online platforms (e.g. Facebook) to remove or label election-related deepfakes and create channels for reporting such content.
        - Candidates and elected officials can seek legal relief if platforms fail to comply with the law. PRIVATE RIGHT OF ACTION
    - Election Deepfakes: AB-2839 addresses the actions of social media USERS who post or repost AI deepfakes that could mislead voters, holding them accountable for spreading false information.
        - There is a MALICE requirement here
        - Does not cover broadcasters
    - Election Deepfakes: AB-2355 requires political advertisements created using AI to include clear disclosures
    - AI and Entertainment: AB-2605 requires studios to get consent from actors before AI'ing their likeness
    - AI and Entertainment: AB-1836 extends similar protections to deceased performers, requiring studios to secure consent from the performers' estates before creating digital replicas
- Vetoed Legislation

        - Regulation of Large AI Systems: SB-1047: Newsome vetoes because it only dealing with "large" systems. Newsom argues that it will give a false sense of security. Asks a task force to find the real and likely harms that might manifest.

- Colorado Artificial Intelligence Act (CAIA)
    - Colorado law targets AI's that are "high risk"
        - Similar in structure to the EU definition of high risk.
        - Formal Definition: High risk constitutes those making consequential decisions relating to:
            - education enrollment or an education opportunity,
            - employment or an employment opportunity,
            - a financial or lending service,
            - an essential government service,
            - health care services,
            - housing,
            - insurance, or
            - a legal service.
        - When deployed "has a material effect on the provision or cost" of one of the above services
    - FOCUS: The goal is to reduce the effect of algorithmic discrimination. More specifically, the concern animating this law is that there could be a material discriminatory effect based on some form of protected class, and the above sectors are particularly important
        - Data might be biased and detection might be harder or masked because the LLM is frequently a black box
    - AMBIT: The imposes obligations on <u>developers</u> and <u>deployers</u> of high-risk AI systems.
        - A "developer" refers to an individual or entity doing business in Colorado that develops or intentionally and substantially modifies a high-risk AI system.
        - A "deployer" refers to an individual or entity doing business in Colorado that deploys a high-risk AI system.
            - Both are required to use reasonable care to protect consumers from any known or reasonably foreseeable risk of algorithmic discrimination arising from the use of high-risk AI systems.
    - ENFORCEMENT: Exclusive to the AG
    - Obligations for Developers (those building or substantial modifying)
        - Documentation requirements: What data has the model been trained on? What are the foreseeable uses and known risks? What data governance measures are in place? Intended outputs. How the system should be used and not used.
        - Disclosure: Mandatory disclosure on the website that they (the public) are using an AI. Documenting mitigation measures (doesn't require that you do them, but demands that you write down what you DID do).
        - Disclosure to the CO AG – if discrimination issues are discovered they must be disclosed
            - INTUITION: Creating a trail of breadcrumbs for assigning liability
            - INTUITION: Creating incentives to create documentation about what happens
            - INTUITION: might have a chilling effect, might compel firms to put their ducks in a row

- Deployers – People who PUT these systems out for use (front office)
    - Notification: Notice to consumers that AI is being used. What consequential decision is being made. Contact information of the firm for customers.
    - When a consequential decision is made using AI, there must be a reason why that it has been made, the type of data that was used must be documented
    - APPEALS: Conditional on an adverse high-risk decision, the deployer must provide
        - The reason for that decision, degree to which the AI contributed to it, and the types of data / sources that were used
        - Consumers have the ability to appeal and correct erroneous data
            - This is a lot like the Fair Credit Reporting Act (FCRA), appeals for denial and reasons for credit
            - This is also a lot like *Louis v SafeRent*
        - There are penalties for not correcting mistakes
    - DISCLOSURES: types of high-risk systems in use that could result in discrimination, how risks are managed (emphasis, discrimination), nature / source of the information leveraged
        - Must disclose actual or likely discrimination to the AG within 90 days if discrimination is discovered. If this is done, you must bring your documentation with you
    - REASONABLE CARE: if an enforcement action is brought, there is a rebuttable presumption that the deployer used reasonable care
        - The law gives you some safe harbor if you have an up to date risk management policy, follow NIST standards, perform impact assessments within 90 days of substantial modification, and are actively monitoring the system regarding discrimination
- Consumers can opt out of their information being used
    - On opt out, it is unclear if firms can penalize for opting out (second degree price discrimination)
- EXEMPTIONS to the law
    - Firms employing fewer than 50 people
    - Firms that do not train the model on their own
    - Firms using a system that is developed for its intended purpose
    - Deployer makes the impact assessment available
        - YOU NEED ALL OF THESE
- AFFIRMATIVE DEFENSES TO AG ACTION: Developers and deployers facing an enforcement action have an affirmative defense if they have:
    - Cured a violation as a result of their own internal reviews or by "red teaming" (i.e., following an internal process to discover risks) or external feedback, AND
    - Complied with the latest version of the NIST AI risk management framework, another nationally or internationally recognized AI risk management framework, or a framework chosen by the attorney general.
        - NEED BOTH!

- Utah Consumer Protection Law
    - SB149 – Unanimous Passage. Colorado is addressing algorithmic discrimination. Utah is addressing consumer protection. The Bill emphasizes that it is not a defense if an AI is the reason a firm violates Utah consumer protection laws
    - Broad Strokes: This bill:

        - (1) specifies that Utah's consumer protection laws apply equally to an entity's use of generative artificial intelligence as they do to the entity's other activities,
        - (2) requires private sector entities to take steps to disclose and/or respond to inquiries about their use of generative artificial intelligence, and
        - (3) creates the Office of Artificial Intelligence Policy which is charged with administering an artificial intelligence learning laboratory program.
    - EFFECTIVE DATE: May 2024
    - DEFINITION OF AI: artificial system that:
        - (i) is trained on data;
        - (ii) interacts with a person using text, audio, or visual communication; and
        - (iii) generates non-scripted outputs similar to outputs created by a human, with limited or no human oversight.
    - AMBIT: Two categories of obligations.
        - Persons (firms) who use prompts etc (deployers)
            - Think Bank of American answering your questions.
        - Person (firms) who provides services or a regulated occupation (deployers)
            - Think of occupations where there is licensing. Runs the gamut from truck drivers to hairdressers
    - ENFORCEMENT: No private right of action. Exclusive to Utah AG and Division of Consumer Protection
    - RELIEF: Among other measures of relief, the Division may impose an administrative fine for up to $2,500 for each violation.
    - REQUIREMENTS:
        - Deployers: Clearly and conspicuously disclose that this is an AI if the person (user) asks if they are interacting with an AI
        - Regulated Occupations: Deployers have affirmative obligation to prominently disclose when a person is interacting with an AI in the provision of regulated services. This must be done verbally at the start of an oral exchange or conversation and through electronic messaging before a written exchange.
    - Statute also creates the Office of Artificial Intelligence Policy. Charged with:
        - Creating and administering an artificial intelligence learning laboratory program
        - Make future regulatory recommendations
        - Invite and receive applications from people to participate in the learning laboratory.
        - A participant who uses or wants to utilize artificial intelligence technology in Utah can apply for regulatory mitigation with the office through a structure created by the law.
    - Statute amends the Utah Consumer Protection Act to state that deidentified data includes synthetic data
    - The KEY is that we apply this to ALL CONSUMER PROTECTION LAW, and then we add some disclosure stuff for the groups
        - INTUITION: You could think about this like the *AF v Character AI* which was brought under Texas Consumer Protection Laws

- New York - Legislative Oversight of Automated Decision-making in Government (LOADinG) Act
    - S7543A, mandates state agencies to assess any software using AI and submit these assessments to the governor and legislative leaders, along with posting them publicly
    - EFFECTIVE DATE: January 2025
    - DEFINITION OF AI: No formal definition of AI.

        - ASSUMPTION: this would seem to mean automated decision making
    - AMBIT: Focus is on state agencies
    - ENFORCEMENT: There is a private right of action for persons harmed
        - There is a presumption that a system created and/or operated in violation of the LOADinG Act caused the harm or violation alleged. Defendant (the state agency) can rebut this presumption with "clear and convincing evidence"
    - REQUIREMENTS: Several,
        - New York state agencies must conduct reviews and publish reports that detail how they're using artificial intelligence software
        - Agencies must perform assessments of any software that uses algorithms, - computational models or AI techniques, and then submit those reviews to the governor / legislative leaders. Reports must also be posted online.
        - Bars the use of AI, automated decision making, for when a citizen receives unemployment benefits or child care assistance, unless the system is being consistently monitored by a human.
        - State workers are shielded from having their hours or job duties limited because of AI under the law.

- Tennessee - Ensuring Likeness, Voice, and Image Security Act of 2024
    - TN is amending the law – gives everyone a property right regarding their likeness. This is amending the ELVIS Act
    - EFFECTIVE DATE: July 2024
    - DEFINITION OF AI: No specific definition of AI.
    - AMBIT: Focus is on persons or entities who knowingly infringe on voice or likeness in any medium. My presumption is that this includes things like deep fakes. Requires publication
        - Carveouts for news, broadcast, criticism, satire, and other standard fare
    - ENFORCEMENT: There is a private right of action and criminal enforcement
        - This is jurisdictionally limited to TN Residence
    - FOCUS: The focus is a change on how we define voice. Must be readily identifiable and attributable to an individual. Can be real or can be a simulation
        - Shift from "photograph or likeness" to "photograph, voice, or likeness".
        - This is similar to NY and LA Statutes
        - The exclusive right to commercial exploitation of the property rights is terminated by proof of the non-use of the name, photograph, voice, or likeness of an individual for commercial purposes by an executor, assignee, heir, or devisee to the use for a period of two (2) years subsequent to the initial period of ten (10) years following the individual's death.

- Michigan Enrolled House Bill No. 5144
    - This is a deep fake bill. Think of the Biden Telephone Calls (*In the Matter of Steve Kramer)*
    - EFFECTIVE DATE: Feb 2024
    - DEFINITION OF AI: No definition of AI. Focus on distribution of deceptive media close to elections. False depictions of an individual
    - AMBIT: Person who distribute or agree to distribute materially deceptive media if:
        - (a) The person knows the media falsely represents a depicted individual.
        - (b) The distribution occurs within 90 days before an election.
        - (c) The person intends the distribution to harm the reputation or electoral prospects of a candidate in an election, and the distribution is reasonably likely to do so

        - (d) The person intends the distribution to change the voting behavior of electors in an election by deceiving the electors
    - ENFORCEMENT: AG, Candidate, Deceived Person, other private parties
    - Carveouts:
        - Does not apply if the fake clearly and persistently indicates it is a fake
        - This differs based on medium of deception (video, voice, image)
    - PENALTIES after injunction for non-frivolous suit
        - First. Misdemeanor. 90 days prison. $500 fine (any assortment thereof)
        - Subsequent violations within 5 years. Felony. 5 years in prison. $1000 fine (any assortment thereof)

- Hawaii Revenge Pornography Law (HRS § 711-1110.9)
    - Statute is about revenge porn and non-consensual private imagery
    - EFFECTIVE DATE: January 2022
    - DEFINITION OF AI: Regulation of use of images, not specific to AI
    - AMBIT: Persons who create or disclosure of fictitious imagery, disclosure or threated to disclose of images of a nude person / one engaged in an intimate act
        - Applies to real images and fakes that look like a real person
    - ENFORCEMENT: Enforcement appears to be exclusive to law enforcement.
    - PENALTIES:
        - A Class C felony is punishable by up to 5 years in prison and a fine of up to $10,000.
        - The court may also order the destruction of any recording made in violation of the law.
    - NOTES: This is about the creation / disclosure of fictitious imagery or real sexual imagery.
        - The image must look like the harmed party
        - Defendant must act with intent to harm (career, finances, reputation, revenge)
        - DOES NOT APPLY to
            - Voluntary commercial transactions (professional pornography)
            - Public acts
        - There is no liability for platforms that the image might be placed on
        - Statute is interesting because it is about what it LOOKS like, not how it is made
        - No carveouts for permissible deep fakes

Federal Regulatory Efforts

- FERRARO taxonomy
    - How is the object of the regulation defined?
        - Is it the thing (what defines AI, 1.21 gigawatts)?
        - Are they concerned with how it is used?
    - What are the exemptions to the regulation?
    - What is required of the parties?
    - Whom is targeted (within the ambit)?
    - Who has enforcement authority (agency, AG, private parties)?
    - Under what authority / what is the justification?
    - By when (timing horizon)?
    - What are the remedies / penalties for non-compliance?
    - What are the exemptions?
    - What are targets of the regulation required to do?

- TAKE IT DOWN Act (Cruz)
    - EFFECTIVE DATE: As of May 2025, passed house and senate. Awaiting approval by the Executive. Enforceable one year from being signed into law.
    - DEFNITION OF AI: Not an AI Regulation. Deals with nonconsensual imagery. Separately deals with minors and those of age. Focus is on proliferation rather than generation
        - A digital forgery is any intimate visual depiction of an identifiable individual. This includes both generative AI and other technological means (e.g. photoshop). Two key parts. The image must
            - Be created by a machine
            - Include an intimate depiction of an identifiable individual. An intimate depiction is any private body parts, sex acts, etc.
    - AMBIT: Establishes Criminal Prohibitions and Obligations of Platforms.
        - See the notes. Focus is on proliferation, not generation
        - Also encompasses threats to disclose actual and fake intimate images
    - ENFORCEMENT: Enforcement by the FTC
    - PENALTIES:
        - Criminal and civil penalties (stronger for minors)
        - Fine or imprisonment (per offence)
        - Two years (three years)
    - EXCEPTIONS:
        - a lawfully authorized investigative, protective, or intelligence activity of a law enforcement agency, an intelligence agency the United States;
        - a disclosure made reasonably and in good faith— to a law enforcement officer or agency; as part of a document production or filing associated with a legal proceeding; as part of medical education, diagnosis, or treatment or for a legitimate medical, scientific, or education purpose;
        - in the reporting of unlawful content or unsolicited or unwelcome conduct or in pursuance of a legal, professional, or other lawful obligation; or
        - to seek support or help with respect to the receipt of an unsolicited intimate visual depiction
        - a person who possesses or publishes an intimate visual depiction of himself or herself engaged in nudity or sexually explicit conduct
    - NOTES: The Act deals in large part with nonconsensual sexual imagery and it treats images of minors and those of age differently.
        - Focus is on the publication or proliferation of the image rather than the generation of those images
        - Authentic imagery and Digital Fakes are the same if the person is identifiable
        - Determinant factor – nonconsensual publication of an intimate image of an adult.
            - Focus is likely here to meet interstate commerce rules
        - INTENT REQUIREMENTS are there – intent to cause harm or actual harm
        - There are rules regarding threats to disclose nonconsensual intimate images
        - This content is forfeited if discovered
        - Rules for platforms – takedown provisions – Notice and removal
            - Within one year the platforms must establish process
            - Platform needs to provide clear and conspicuous notice of process
            - Within 48 hours of notification, the image and similar copies be removed
                - Platform is not liable if they make a good faith effort to remove

        - Similar to the DMCA takedown good faith effort

- Hawley Decoupling Act
    - I am not going to waste time on Hawley's nonsense and a formal prep. This thing will never pass. Basically, Hawley wants to split all AI related work, research, exchange, and everything else with China, Chinese nationals, Chinese firms etc etc etc.
    - No import, no export, no collaboration.

- Biden EO 14110
    - EFFECTIVE DATE: the EO has been rescinded by the Trump Administration
    - DEFNITION OF AI: There is a 1.21 gigawatts definition in here that is already out of date
        - any model that was trained using a quantity of computing power greater than 1026 integer or floating-point operations, or using primarily biological sequence data and using a quantity of computing power greater than 1023 integer or floating-point operations; AND
        - any computing cluster that has a set of machines physically co-located in a single datacenter, transitively connected by data center networking of over 100 Gbit/s, and having a theoretical maximum computing capacity of 1020 integer or floating-point operations per second for training AI.
    - AMBIT: The focus of this section is ensuring safe and reliable AI. There is an understandable focus on dual-use models which pervades Biden's EOs. See Notes
        - Companies developing or demonstrating an intent to develop potential dual-use foundation models to provide the Federal Government, on an ongoing basis, with information, reports, or records (see reporting below)
        - Companies, individuals, or other organizations or entities that acquire, develop, or possess a potential large-scale computing cluster to report any such acquisition, development, or possession, including the existence and location of these clusters and the amount of total computing power available in each cluster.
            - This is an AWS flag
    - ENFORCEMENT: Variety of national intelligence agencies
        - Pursuant to the Defense Production Act
    - PENALTIES: Shame
    - EXCEPTIONS: This is for ALL dual use model creators and Infrastructure as a service people who meet the thresholds (see above)
    - NOTES:
        - The term "dual-use foundation model" means an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by:
            - (i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons;
            - (ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber attacks; or
            - (iii) permitting the evasion of human control or oversight through means of deception or obfuscation.

        - THESE ARE EXAMPLES – Umbrella – posing a risk to human safety, national security
        - Exhibits or could be modified to exhibit high performance at tasks. This is a wide definition
        - Even if there are technical safeguards, if it could be used, it still qualifies
    - Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities. Within 90 days, secretary of commerce, requires
    - If you are working with such models, you have reporting requirements. What are the reports you need to produce?
        - (A) any ongoing or planned activities related to training, developing, or producing dual-use foundation models, including the physical and cybersecurity protections taken to assure the integrity of that training process against sophisticated threats;
        - (B) the ownership and possession of the model weights of any dual-use foundation models, and the physical and cybersecurity measures taken to protect those model weights; and
            - JAFFER makes the not unreasonable point that model weighting is black boxed and evolves making this nearly impossible. Also if this is a security issue, why would you have people copy them and hand them over to the feds when they could also be hacked. The firm is just as incentivized to keep it safe and arguably more skilled
        - (C) the results of any developed dual-use foundation model's performance in relevant AI red-team testing based on guidance developed by NIST pursuant to subsection 4.1(a)(ii) of this section, and a description of any associated measures the company has taken to meet safety objective
            - JAFFER points out some flaws. Are you red-teaming, you don't have to? Is this a shaming tool to compel red teaming? In short, people who aren't red-teaming enough become de facto regulated. I see things a little closer to FERRARO. Wide guidance. Formal regulation comes with failures
    - Also deals with Infrastructure as a Service (IaaS) (i.e. cloud computing services), e.g. (AWS, Azure). Requirements on reporting the existence and use of clusters.
        - Econ argument, this will increase costs
        - Brad response, if you have a billing system registration is already there. De minimis chilling effect
- PROFESSOR COMMENTS
    - The intent is safeguarding the safety, effectiveness, and reliability of AI.
    - FERRARO emphasizes that regulation is not 0/1. He views the approach as a gentler touch that straight regs.
        - Bears note that the current administration is turning this off. Just playing with different dials
    - Order Was: Voluntary Commitments from large firms → Then the EO based on commitments that were then applied to others →
    - On weights – Model weights were thought of at the time as very important and proprietary. Now they are more loosely being shared.

- Trump Revision to EO 14110
    - Repeals the Biden EO almost in entirety
    - Within 180 days, the administration will convene a committee and promulgate new rules
    - The goal: protect AI and save the world, establish American dominance. Text is thin on details and high on goals. Its more or less a reset on Biden while heading in the same general direction
        - Notes: A new rule is coming – gotten nearly 9000 comments already

- Vance Commentary from France
    - Not a statute. Vice Presidential comments in France. Best indicator of current (May 2025) administration position. Four main points that are rather high minded but non-specific
    - 1) Ensure American AI is the gold standard of AI – partner with the world
    - 2) Do not kill AI emergence with excessive regulation
        - I read this as a pot shot at GDPR
        - He is also concerned with moat-building for incumbents, not invalid
    - 3) Free from ideological bias and not co-opted by authoritarianism
        - I read this is as free from anti-conservative bias lest the Little Mermaid be Black. Standard talking point for Vance
    - 4) Pro-worker path - Education and uplifting workers. Retraining for the new economy
    - This raises the important point of Federal pre-emption. Vance doesn't bring this up, the class does. Active debate. Firms might want this to avoid dealing with conflicting regulatory regimes. States might want it as well to put it on the Feds. States might also not think this is federal purview.
    - Two Types of pre-emption
        - Field Pre-emption - In order to pre-empt state law in its entirety, the federal government has to be clear that it is **attempting to "occupy the field" under article I**
        - Conflict Pre-emption – individual level conflict when it is unclear that the govt is attempting to occupy the field or is building something not clearly under an article I power

European Union (EU) AI Act

- Surface level notes
    - The EU AI act is impenetrable. So, we are focusing on the summaries because they can actually be
    - Thing v Use – this is a taxonomy FERRARO has put forward before.
        - Are we thinking about the AI as a technological artifact or are we focusing on use
        - The EU Regulation is primarily focused on Use, but there are elements of "Thing"
        - This is a lot like Arizona's law in healthcare, publication of intimate imagery
        - When the EU is discussing the THING – back to Dual Use Foundational Models
    - FERRARO notes that when the EU started, the bloc focused almost exclusively on USE.
        - When ChatGPT dropped, a new section on Generative AI got tacked on (EU calls it General Purpose AI). This is a completely different framework
        - There has also been some push back from the French and the Germans as their own Generative AI firms have begun to spin up (Open AI competitors in nascency)

- DEFINITION OF AI
    - Recall, the base push at the EU was on use, but there are definitions of the thing

    - GAI Basic – "General-purpose AI model' means an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications. This does not cover AI models that are used before release on the market for research, development and prototyping activities.
        - Technical Definition - 10^25 FLOPS. Catch all level to meet General Purpose AI
    - GAI Systemic Risk - A risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the internal market due to its reach, and with actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain

- ENFORCEMENT TIMELINE (From August 2024)
    - 6 months for prohibited AI systems
    - 12 months for GPAI.
    - 24 months for high risk AI systems under Annex III (listed below).
    - 36 months for high risk AI systems under Annex I (other high risk systems not specified)
    - Codes of practice must be ready 9 months after entry into force.

- Obligations based on type of AI Use

	Examples	Provider Obligation aka Developers	Deployer Obligation	Notes
Prohibited	Social Scoring, Manipulative AI, Citizen monitoring for "scores", anything in Minority Report	Can't be done - prohibited	Can't be done - prohibited	
High Risk	Almost everything. There are a lot of carveouts here (Annex III below). High risk is anything that can yield harm to health, safety, or fundamental rights. This includes non-banned biometrics, critical infrastructure, employment, financial decisions, education, administration of justice and the democratic process. FERRARO points out that you could use AI to predict train-track failure. The core issue is whether the decision maker is being PROSCRIPTIVE or DIAGNOSTIC. There are exceptions for	Establish risk management systems, establish data governance rules, technical documentation, ensure transparency to users that they are engaging with AI, record keeping, instructions for use for downstream developers, human oversight, compliance with standards for security and accuracy, registration of the system with the EU, third party conformity assessments, immediately take corrective action on	Complete a Fundamental Rights assessment if the AI is touching on important areas (finance, health, employment, critical infrastructure), implement human oversight, use only relevant data, suspend the system if a fundamental risk emerges, reporting to the provider (developer) of the AI system if issues emerge, compliance with GDPR / AI Act, inform people they are using a high risk AI system (disclosure)	Note the balance that this doesn't apply to national military apparatus (the French could make killbots) but does apply to civil law enforcement. Differentiated national sovereignty issues

	narrow procedural tasks that have been done by a human before as long as the human is not removed from the loop (very fact dependent)	observed failure, monitoring performance and safety		
Minimal Risk	Spam Filters, Chat Bots, video game systems	Design ways to alert people that they are using an AI system (disclosure)	Attain consent from users, disclose and label "deep faked" content when provisioned to users	

- Annex III Carveouts
    - Systems are not high risk if they do one of the following (modifiable by the Commission)
        - the AI system performs a narrow procedural task;
        - improves the result of a previously completed human activity;
        - detects decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment without proper human review; or
        - performs a preparatory task to an assessment relevant for the purpose of the use cases listed in Annex III.

- Annex III Use Cases (comprehensive from report)
    - These are all high-risk, but they are not banned
    - Non-banned biometrics:
        - Remote biometric identification systems, excluding biometric verification that confirm a person is who they claim to be.
        - Biometric categorization systems inferring sensitive or protected attributes or characteristics.
        - Emotion recognition systems.
    - Critical infrastructure:
        - Safety components in the management and operation of critical digital infrastructure, road traffic, and the supply of water, gas, heating and electricity.
    - Education and vocational training:
        - AI systems determining access, admission or assignment to educational and vocational training institutions at all levels.
        - Evaluating learning outcomes, including those used to steer the student's learning process.
        - Assessing the appropriate level of education for an individual.
        - Monitoring and detecting prohibited student behavior during tests.
    - Employment, workers management and access to self-employment:
        - AI systems used for recruitment or selection, particularly targeted job ads, analyzing and filtering applications, and evaluating candidates.
        - Promotion and termination of contracts, allocating tasks based on personality traits or characteristics and behavior, and monitoring and evaluating performance.
    - Access to and enjoyment of essential public and private services:
        - AI systems used by public authorities for assessing eligibility to benefits and services, including their allocation, reduction, revocation, or recovery.
        - Evaluating creditworthiness, except when detecting financial fraud.

        - Evaluating and classifying emergency calls, including dispatch prioritizing of police, firefighters, medical aid and urgent patient triage services.
        - Risk assessments and pricing in health and life insurance.
    - Law enforcement:
        - AI systems used to assess an individual's risk of becoming a crime victim.
        - Polygraphs.
        - Evaluating evidence reliability during criminal investigations or prosecutions.
        - Assessing an individual's risk of offending or re-offending not solely based on profiling or assessing personality traits or past criminal behavior.
        - Profiling during criminal detections, investigations or prosecutions.
    - Migration, asylum and border control management:
        - Polygraphs.
        - Assessments of irregular migration or health risks.
        - Examination of applications for asylum, visa and residence permits, and associated complaints related to eligibility.
        - Detecting, recognizing or identifying individuals, except verifying travel documents.
    - Administration of justice and democratic processes:
        - AI systems used in researching and interpreting facts and applying the law to concrete facts or used in alternative dispute resolution.
        - Influencing elections and referenda outcomes or voting behavior, excluding outputs that do not directly interact with people, like tools used to organize, optimize and structure political campaigns.

- PROVIDER OBLIGATIONS (DEVELOPERS)
    - Keep and maintain technical documentation, make it available, respect copyright law, disseminate details on what has been used for training,
        - There is a carveout for free and open source models
    - Additional obligations for Systemic Risk must provide model evaluations, assess and mitigate risks, document and report serious incidents, conduct red-term adversarial training, ensure sufficient security
        - A risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the internal market due to its reach, and with actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain

- DEPLOYER OBLIGATIONS
    - These are listed in the chart. Primarily rights assessment, notification of issues upstream, etc

- REMEDIES
    - These are largely fines
    - Top Tier – €35mm or 7% of global revenue
    - In the case of small and medium enterprises, fines will be as described above, but whichever amount is lower.
    - National competent authorities will determine the fines in line with the guidance below

Noncompliance case	Proposed fine
Breach of AI Act prohibitions	Fines up to €35 million or 7% of total worldwide annual turnover (revenue), whichever is higher
Noncompliance with the obligations set out for providers of high-risk AI systems or GPAI models, authorized representatives, importers, distributors, users or notified bodies	Fines up to €15 million or 3% of total worldwide annual turnover (revenue), whichever is higher
Supply of incorrect or misleading information to the notified bodies or national competent authorities in reply to a request	Fines up to €7.5 million or 1% of total worldwide annual turnover (revenue), whichever is higher

- REGULATORY ENFORCEMENT
    - At an EU level, the AI Act governance framework also establishes the:
    - AI Office**:** within the EU Commission, but with functional independence This new body will have oversight responsibilities for GPAI models. It will contribute to the development of standards and testing practices, coordinate with the national competent authorities and help enforce the rules in Member States
    - AI Board**:** representing the Member States to provide strategic oversight for the AI Office The Board will support the implementation of the AI Act and regulations promulgated pursuant to it, including the design of codes of practice for GPAI models
    - Scientific panel of independent experts**:** to support the activities of the AI Office The panel will contribute to the development of methodologies for evaluating the capabilities of GPAI models and their subsequent classification, while also monitoring possible safety risks
    - Advisory forum**:** with representatives of industry and civil society Will provide technical expertise to the AI Board

Voluntary Commitments

There six different commitments organized below based on timeline, from Biden, to Hiroshima, to Bletchley, to the UN. Paladin and DIANA are "private" discussions from firms about the commitments they are taking on.

In class discussion is vague about how these sorts of things will be used. One theory is that this is a flag being thrown to industry to self-regulate so the govt doesn't need to regulate. Others theorize this as a shot across the bow. If you botch these, there will be oversight, because this is how we make end up with all laws. Another theory that this is a stopgap before the Biden EO, and it got the CEOs into meetings with people on the Hill. The initial voluntary commitments were among small set of firms (<10) and the EO was meant to expand them to cover everyone playing in the Gen AI Space.

- White House Voluntary from Biden (2023)
    - These are voluntary commitments from large Gen AI firms prior to any formal regulation
    - SAFETY
        - Committing to red teaming (having dedicated internal and external teams attempting to uncover the vulnerabilities of systems)
        - Assess bias and discrimination, cyber threats, biological chemical and radiological, self-replication (Skynet)
        - Work towards information sharing across public and private agents to promote trust, share safety risks, disseminate new capabilities, and highlight new risks (i.e. establishment of standards and mutual understanding)
            - Reducing information asymmetries across groups

        - Sharing standards (e.g. NIST)
    - SECURITY
        - Invest in cyber security to protect model weights against insider threats
            - Cyber and physical security
            - NOTE: This is back in the day of closed models. Model weight security is less of an issue now
        - Incentivize third party discovery of issues and vulnerabilities
            - Bug bounties, contests, prizes (standard white hat hacking)
    - TRUST
        - Develop and deploy mechanisms that enable users to understand that content is AI generated
            - Flags provenance, watermarking, audio and visual
            - Develop tools to detect if such things were created by their system
        - Embedding meta-data watermarking
        - Publicly report model and system capabilities, limitations, domains of appropriate and inappropriate use, and risks
            - Published reports, etc. Limitations,
        - Prioritize research on societal risks of AI systems, including avoiding harmful bias and discrimination, and protecting privacy
        - Develop and deploy frontier AI systems to help address societal grand challenges. AI for GOOD for ALL!

- Hiroshima Process (G7 Nations) – Oct 2023
    - Drops the same day as the Biden Executive order (EO 14110)
    - The Hiroshima AI Process is a G7 initiative aiming to promote safe, secure, and trustworthy AI worldwide. Code of conduct for developing advanced AI systems
        - Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States
    - Take appropriate measures when developing advanced AI systems to identify risks pre-deployment (testing) and across the AI development lifecycle
        - Risk, bias, the whole shebang
    - Identifying and mitigating vulnerabilities, as well as patterns of misuse after deployment
    - Notify and report to the public limitations
        - make information sufficiently transparent for use by users, developers, and deployers
    - Information sharing and reporting incidents across industry, government, academia, and civil society
        - This should safeguard intellectual property rights (trade secrets)
    - Develop, implement, and disclose AI governance policies and risk management practice
        - Safeguarding PII, be clear on the process
    - Invest in robust cyber security and physical security and safeguard against insider risk
    - Develop and deploy reliable content authentication where technically feasible
        - This is flagging AI generated content
    - Prioritize research to mitigate societal safety risks and prioritize investment in effective mitigation measures
    - Prioritize investment in AI systems to address societal grand challenges – AI for GOOD
        - Includes climate, health, education, etc.
    - Advance the development and adoption of international technical standards

    - Implement appropriate data input measures and protections for personal data and intellectual property
        - Emphasis on data input measures for PII and IP
        - This is distinct from the others

- Bletchley Declaration – Nov 2023
    - Bletchley Park is where they broke ENIGMA. Somewhat fitting that it emerges here.
    - This is a broader group of countries. From Rwanada and Singapore to the Saudis to the EU to the Commonwealth of Nations. Global south, middle east, whole nine yards
    - Focus is on frontier AI models – examine and address (attention) on AI risks, harms of foundation models. The concerns focus on cybersecurity, disinformation, bio risks, etc.
    - COMMITMENTS
        - Resolve to work together to develop AI that is safe, measure and mitigate threats, and cooperate at the national and international level.
        - Building risk based policies together
        - Human centric approach (AI for social good)
    - This is very kumbaya. We're all in it together. We'll figure out this AI thing together

- UN Resolution – March 2024
    - First UN resolution on the matter. Lots of throat clearing as all UN resolutions enjoy a good amount of throat clearing and self-congratulations. Now we basically have all the countries
        - Focuses on non-military uses (similar to the EU AI act)
        - Focuses on it being human centered (similar to Bletchley)
        - Focuses on AI human harms being bad
    - CONTENT
        - Bridge the digital divide – transferring technology to close this divide (slightly new but emergent in Bletchley)
        - Respect human rights
        - Develop safe, secure, and trustworthy AI systems
        - Calls on nations to cooperate with each other
        - Content authentication
        - Protecting IP
        - Training and testing
        - Linguistic and cultural diversity (this is new)
        - Closing the gender divide (new)
        - Promoted information sharing among entities across the lifecycle of AI development and deployment lifecycle
        - Impact on labor markets (this is new)
        - We govern AI, it will not govern us – Anti-Skynet provision
    - Caveats are everywhere but not explicated in the article

- Paladin Investment Principals – 2024
    - JAFFER is a venture partner with these guys. Capital investment group in DC. Note that we are moving into the private sector. These are the voluntary commitments that the venture partners want investees to make in order to receive capital. Companies that don't follow these wont get investment dollars.
    - SECURITY
        - Ensure the companies they invest in are taking steps to protect themselves against cyber attacks

        - Identify and mitigate risks by investing in software that is secure-by-design and resilient-by-design (security responsibility is shifted from the end user to the developer)
        - Ensure companies are taking steps to mitigate risk in the software supply chain
        - Ensure firms are incentivized to discover and report bugs, and engage in rapid remediation
    - TRUST
        - Adopting industry standards and best practices to protect security and privacy of customer data
        - Ensure firms only sell to countries that abide by international law as recognized by the US
        - Require the companies to adhere to US sanctions and relevant regulations (export controls)
        - Ensure firms we invest in identify trust and safety solutions, identify risks with them, take steps to mitigate them, and share that information with appropriate entities
        - Ensure we know our co-investors and avoid co-investment with investors that are in countries of concern to the United States (know your partners)
    - SAFETY
        - Red-teaming – investment in such behavior pre and post deployment
        - Ensure companies we invest in engage in robust testing to protect against misuse, national security concerns, and consider human alternatives to fully automated systems
        - Ensure companies invest in appropriate cybersecurity and AI best practices (e.g NIST)
    - NATIONAL SECURITY
        - Invest in companies that will enhance national defense, national security, foreign policy objectives etc that protect critical society and a free world
        - Avoid investments that would undermine the defense and national security and foreign policy interests of free and open societies

- Allied Capital Community (ACC / DIANA) – 2024
    - The ACC is a network of investors from NATO member nations who are aligned with NATO's values and mission, particularly focusing on emerging and disruptive technologies. The ACC, established by NATO's Defence Innovation Accelerator for the North Atlantic (DIANA), aims to support and accelerate the development of dual-use technology solutions that can benefit both civilian and military applications
        - Must be in a NATO member nation to get these commitments
    - CORE ISSUES
        - Strategic alignment – uphold democratic values and principles that undergird them, and will consider at the time the impact on free society
        - Principles of responsible use – focus on the trust, safety, and societal risk (secure-by-design and resilient-by-design) and take affirmative action to mitigate risk
        - Security – comply with applicable laws, protect the security and privacy of PII data, taking affirmative steps to identify and mitigate risks in their supply chains,

Patenting Cases

Thaler v Vidal

- FACTS:
  - Thaler creates an AI program that can spontaneously generate designs and art. DABUS - Device for the Autonomous Bootstrapping of Unified Sentience
  - Because patents require an "individual" to be an inventor, Thaler submitted design on DABUS's behalf. Thaler lists DABUS as the inventor. He attached three things
    - Assigns the rights of the patent himself (Thaler)
    - Submits the sworn oath on DABUS's behalf
    - Supplement explaining that DABUS is a creativity machine
  - USPTO denies the patent application. Thaler appeals to the 4th Circuit (judicial review) under the APA. File cross motions for summary judgment
  - District court says DABUS is not an individual (*35 USC 101*). Thus, can't be an inventor
    - "whoever invents a new or useful process"
  - Federal Circuit applies *de novo* review – full start. No reference to the prior courts finding
  - Pertinent law standard – Administrative Procedure Act - Arbitrary and capricious application
- ISSUE: Can an AI be an individual?
  - NO!!!!
- HOLDING
  - There is no ambiguity in the statute, so the review of the text is the beginning and end of the analysis
  - AI is not an individual, so it cannot be an inventor
- REASONING
  - Any natural interpretation of the term individual means a person
  - The Act doesn't define what an individual is
    - But the dictionary does
    - Common usage does
- THALER's Argument
  - Natural reading doesn't consider the true meaning of the law. Supporting innovation
    - Court says no. We look to the law. The statute says what the statute says
  - Thaler says the act says "whoever"
    - Court says no. Read the whole act. The US Code sets up the fact that it needs to be an individual. i.e. person
  - Thaler says there is a constitutional avoidance issue
    - Court says no. Congress has acted. So, we cannot argue that this is unconstitutional because Congress is acting on a Constitutional grant of authority
  - Thaler says "what about South Africa"
    - Court says great, move to South Africa

Copyright Cases

Warhol Foundation v Goldsmith

- FACTS:
  - 1981 – Goldsmith (photographer) takes a picture of Prince. Licenses that to Vanity Fair for one use, no other uses (cover and a half page). Vanity Fair asks Andy Warhol to mess with it, makes a bunch of silk screen prints (different colors). This is known as the Prince Series. The orange one is used as a cover on the magazine in the 1980s

    - When Prince dies, Conte Nast wants to do a memorial, contacts the Andy Warhol foundation, they use the Purple Picture which is published. Goldsmith sees the picture, contacts AWF about potential violation. AWF sues for declaratory judgement.
    - District court finds that the work was sufficiently transformative, so there is fair use. Warhol is a new expression. Goldsmith is just a photographer
    - Circuit reverses – insufficiently transformative. Changing the color isn't transformative. It has to have "fundamentally new" character. CERT
- ISSUE: is this sufficiently transformative?
    - SOTOMAYOR: NO! This is not sufficiently transformative
        - Takes a very formalistic approach. You look at this side by side. "Share substantially" the same purpose, "no new meaning or message" as a commentary on celebrity, has the "purpose and character" of the use change (both are magazine covers) – maybe you could put it in a museum
- GORSUCH Concurrence:
    - Focuses on the subject object distinction
    - Argues that this is not subject to a parody and satire distinction
        - Potter Stewart test, know it when you see it
    - Use is commercialism. The magazine is reusing the cover. If this was educational or in a museum we might see it different.
- KAGAN Dissent: Yes, it is transformative.
    - Andy Warhol did this all the time
    - Dramatically changed the aesthetic
    - The Court has cited Warhol as transformative based on his goofy art anyway
    - Conte Nast valued the picture being purple because it evoked different emotion
    - KAGAN discusses the mechanical process of building the screens (labor intensive)
        - Warhol is a pioneer here.
- INTERPRETATION: Formalism (majority) v expressionism (dissent)
    - SOTOMAYOR and GORSUCH are interested in the use of something. Has it really changed.
    - KAGAN: more about the narrative content.
    - The criticism becomes that you force judges to be art critics
        - Authors of the law review argue that art is subjective and that's the whole point
        - You are weakening Fair Use protections
- WHY is this all relevant to AI?
    - There are core ambiguities which constitute transformative use, and the way AI is trained is that it takes a bunch of protected stuff and beep boops it into a new thing

Zarya of the Dawn

- Clarity issue: When AI Models train, the first things they do is copy data wholesale. This is a core issue because it is data reproduction. This is an important baseline
- FACTS:
    - Kashtanova writes a graphic novel which is originally granted copyright. It is revealed that Midjourney was used during the production of the images. USCO requests further information about how Midjourney was used.
    - Through Counsel, Kashtanova responds that she:
        - 1) used it as a tool the same way a photographer would use a camera (no different than Photoshop)

        - 2) used many prompts to curate a copyrightable compilation of many many pictures that Midjourney generated in building the graphic novel
    - USCO splits the baby. Says she is the author. Owns the text, arrangement, characters, etc
        - She cannot copyright the images.
        - The mere repetitive prompting of the tool does not change the fact that it is a tool because she has no control over what she is getting
            - This is like hiring another artist, not like using a camera
- HOLDING
    - Somewhat narrow, if there was more control, maybe this would be copyrightable. The issue is randomness, and because it isn't your own mental conception, it isn't copyrightable. Textual prompts are suggestions
        - FERRARO agrees this is narrow. In the film industry there is postproduction and that work permits you to copyright what comes initially out of an AI
        - Edits and the ARTIST doing an editing is a key issue
- GUIDANCE Coming out of this:
    - Autonomously generated content absent human involvement cannot be copyrighted
        - This is based on the understanding of Gen AI that is currently available. The system produces random images that the artist doesn't have CONTROL
        - CRITICAL: based on current technology
    - SUBSTANTIAL HUMAN INVOLVEMENT
        - Does the ordering matter (Brad terrible artist hypo). "Take specific action to my own conception rather than". The stochastic process of the AI needs to be corrected by the artist in some way

IP Infringement Cases

NY Times v Microsoft and Open AI

- PARTIES
    - NY Times – Mainstay of traditional journalism for nearly 200 years
    - Open AI et al – emergent GPT producer
    - Microsoft – We all know who Microsoft is
- FACTS
    - NY Times is files suit against Open AI et al. The relationship between Microsoft and Open AI is a financial and technical partnership where MS provided "seed" funding and then will have a vested stake. MS also provides technical infrastructure. MS also has access to underlying data associated with training the LLM and was engaged with the fine tuning.
    - The allegation is that Microsoft and Open AI are using copyrighted information in commercial products in a way that doesn't comport with Fair Use. This was achieved through the training of the LLM using CommonCrawl, WebText (Reddit), WebText2 (Reddit), Wikipedia and various other "scraped" repositories which included NY Times material. Open AI had not licensed the NY Times Material
    - Important Detail: The Open AI project was a non-profit entity that was backed by large dollars which has transitioned into a for profit. Non-profit is important as it goes to the heart of whether the reproduced material was used in a "transformative way", a "competitive way", if the use was commercial or educational, etc
        - RECALL: The use of the copyrighted material is what matters although the commercial nature of the enterprise is informative. A for-profit entity might not use something commercially. And a not-for-profit entity might use things commercially

- MEANS OF HARM (SUBSTANCE OF THE CLAIM)
    - Unauthorized reproduction of Times works during TRAINING
    - Unauthorized reproduction of Times works
    - Unauthorized attribution of Hallucinated content attributed to the Times
    - Wirecutter review scoops – Wirecutter makes money on click through links on the website
        - The economic theory underpinning this is almost purely freeriding. The exception is the Hallucination arguments
        - Reduces incentives for subscriptions
        - Forecloses Wirecutter clicks
        - Reduces revenues through ad display
        - Tarnishes reputation through hallucination
    - OPEN AI RESPONSE
        - That wasn't our intent, anyone who is doing it is violating our terms of service
        - NY Times cannot copyright the FACTS
        - These are cherry picked examples
        - The harm is speculative, we tell people to go to the Times
        - We are doing much more than just these things
        - We couldn't have possibly foreseen this. There was no intent
    - Best NY Times response
        - A little theft is still theft
        - The issue is not the theft of facts, it is about the unique work and the body of how it is displayed
        - YOU have no evidence that this is not happening on a large scale
        - If you didn't think this would happen, you obviously weren't paying attention
        - The Code of expression (NY Times Style) is a critical part of what we do. There was deliberate attempt to copy the style which goes to the core of the service we provide
        - There is also the training issue, unauthorized use of the data to train the models
    - The Claim of Tarnish through Hallucination
        - Open AI claims this is a training issue. This is an overfitting issue. This is because AI won't say "I don't know"
        - Best Response: Doesn't matter, you are still tarnishing our reputation
    - JAFFER Asks the question, is this a heads I win, tails you lose? When it reproduces exactly what we said, that's infringement? When it hallucinates, that's harming us?
        - I think that's an incorrect backout argument. The harm manifests either way.
- ISSUE
    - Does the training of Chat GPT using copyrighted materials constitute fair use?
        - We don't have a full response from a court on this, but many of the claims survived motion to dismiss. We do not know if summary judgment will be granted.
        - Claims against Microsoft did not survive motion to dismiss.
- MOTION TO DISMISS OUTCOME
    - Case is now aggregated. Combines the complaints from NY Times, Daily News, and the Center for Investigative Reporting (Reveal News and Mother Jones)
    - Statute of limitations issue
        - Open AI claims statute of limitations (3 years), which we are after, applies. The court is looking at when the Times became aware, or should have become aware in the exercise of reasonable diligence.

        - Court dismisses this because Open AI fails to articulate why their behavior, if known to the plaintiff, should have put them on notice that there was infringement. *Cf. McGlynn.*
    - Insufficient Evidence
        - There were NYT article about Open AI reading the internet in total because the articles being cited don't indicate that the NYTs data is being used.
        - The articles in the Times itself are insufficient because although there are trillions of words of the internet, but the article doesn't say the NYTs words are in those trillions in particular
    - Open AI makes a sophisticated publisher argument
        - Sophisticated publisher – heightened expectation of plaintiff
        - Second Circuit has rejected this argument. Doesn't apply here (maybe in 9)
        - This could come out in discovery
- Central claim: direct copy right infringement - Survives
    - First direct infringement claim: Training the model and then disseminating. Allegation is that the data is copied and then permanently stored in memory without permission. Open AI did this using direct scrapes, links posted on reddit, (WebText and WebText2) which were built by OPEN AI. They are taking copyrighted material without permission and using it to build their model
    - Weakest part of this claim?
        - Anyone can do this! They didn't break into the NYTs headquarters.
    - Second direct infringement claim: After copying protected material, Open AI encoded it in a numerical format, trained their model against the stored information, and modified weights until ChatGPT could reproduce the NY Times Style
        - BACKUP CLAIM: if it messes up what the NYT says then the hallucination undermines the NYTs credibility
- Contributory infringement claim
    - Claim is that plaintiffs contributed to infringement by end users by:
        - (1) building and training their LLMs using plaintiffs' works;
        - (2) deciding what content is outputted by their LLMs through specific training techniques; and
        - (3) developing LLMs capable of distributing copies of plaintiffs' works to end users without authorization by plaintiffs.
- Applicable standard
    - (1) direct infringement by a third party,
        - See above
    - (2) that the defendant had 'knowledge of the infringing activity,'
        - Plaintiffs claim is that actual or constructive knowledge; namely, whether defendants objectively "know or have reason to know" of the direct infringement by third-party end users. That is the standard in the Second Circuit. *Gershwin Publ'g Corp*
        - Defendants urge a heightened standard, contending that liability for contributory copyright infringement requires that the defendant have possessed actual knowledge of or willful blindness to specific acts of

infringement. That is the standard in the Ninth Circuit. See *Ludvarts, LLC v. AT&T Mobility*

- We're in the second circuit so who cares. Bound by 2nd Circuit

- Court says we meet this requirement because there are widely publicized incidents and hundreds of examples.
    - Do they have actual or constructive knowledge? Yes, based on statements that the NYTimes told them. And this was core business model. They were training the model specifically, and adjusting them, to ensure that thye were regurgitating the actual text of the Times. They WERE TRYING TO WHILE ON NOTICE. That's the whole idea

- (3) and that the defendant 'materially contribute[d] to' the third party's infringement." *Dow Jones & Co., Inc. v. Juwai Ltd*
    - Standard: To establish that the defendant "materially contributed" to the infringement, the complaint must show that the defendant "encouraged or assisted others' infringement[] or provided machinery or goods that facilitated infringement." *Arista Recs. LLC v. Lime*
    - Court says yes they they have. They have directly provided goods to do so. Differentiation with *Sony* and *Grokster*

Thompson Reuters v ROSS

- BACKGROUND: Judge BIBAS is sitting for this one. FERRARO likes him. Young guy. Son of Greek immigrants. Apparently, he is a savant. Yale. AUSA in NY. Prosecuted a bunch of people
    - Sitting by designation is a swapping thing between appellate and district level
- FACTS
    - ROSS is trying to develop a commercial tool to train on WestLaw. The whole idea is that you could feed the Key system from WestLaw into an AI and have that reproduce the product. Thompson says no. ROSS subsequently uses bulk memos which are largely derived from WestLaw headnotes
    - Allegation is that the use of the Bulk memos is a violation of fair use.
- STANDARD AND RESULT
    - Court shall grant summary judgment if there is no material dispute as to the facts, as a basis of law, how does it come out. It is a question of law
    - Two elements of fair use doctrine weigh in Thompson Reuters favor.
    - They win as a matter of law
- ANALYSIS
    - Four-part analysis of fair use doctrine for Direct Copyright Infringement
        - First, the court found that Thomson Reuters owned valid copyrights, although it allowed disputes as to the originality of specific headnotes to go to trial
        - Second, the court analyzed whether Thompson Reuters had shown actual copying of Westlaw headnotes and substantial similarity between the headnotes and the Bulk Memos used as training data. There is sufficient evidence for several
        - Finally, the court gave short shrift to ROSS's defenses of innocent infringement, copyright misuse, merger, and scenes à faire
- Fair Use Analysis
    - Purpose and Character – We look at *Warhol*. Both parties are trying to do the same thing. They are creating a legal research tool. Advantage Thompson Reuters.

    - Nature of the Copyrighted Work – Headnotes from WestLaw are creative, but they aren't a novel. They aren't creative enough. Advantage ROSS
    - Amount and substantiality of use - "[w]hat matters is not 'the amount and substantiality of the portion used *in making a copy*, but rather the amount and substantiality of *what is thereby made accessible* to a public for which it may serve as a competing substitute.'" Advantage ROSS
        - FERRARO opines that making thing accessible is an important distinction in Gen AI because there is a lot of black boxing
    - Effect of the Use on the Potential Market – The intent is to compete directly. This clearly will have an effect on the potential market. Advantage Thompson Reuters
- OUTCOME: We now proceed to trial. FAIR USE NOT A DEFENSE. Proceed to trial on damages and whether the copyrights are viable.

Cases Discussing Other Forms of Liability

AF v Character AI – Product Liability

- PARTIES
    - AF – Minor children and their parents.
    - Character AI – online chatbot that allows people to chat with a character. Underwritten by Google. Interactive thing that lets you chat with people like Danerys Targaryen
- FACTS
    - Character AI developed an LLM. Google is being sued because the founders are previous Google employees, Google provided the training data, and underwrote a lot of the project
    - Allegation is that the product is developed by former Google employees which mimics personalities. User can input parameters. You can also describe your own characters and then you can chat with them
    - Additional claim is that children are targeted. In the fact pattern, a child claims that they are not getting enough screen time. The AI tacitly encourages him to kill their parents
    - Parents are suing under a series of liability and product liability claims
        - Marketed to children and encourages them (children) to engage in sexual or violent behavior (strict liability)
        - Negligence claim – failing to account for the behavior that would obviously happen
    - Other claims
        - Engage in anorexia – eliminating food intake, advice on how to engage in illegal behavior (embezzlement), adult predators were given a space to reveal their abuse of children, practicing psychology on children, COPPA violation, collecting data on children, limiting parental consent
    - LEGAL CLAIMS
        - Strict liability – product is inherently dangerous. Defective design. Harms were foreseeable, could have been mitigated, and those steps were not taken. "If used as intended"
        - Strict liability – failure to inform people of the risks. (Adequate warning to consumers and parents about foreseeable risks of mental and physical harm). Foreseeable danger, that others wouldn't have knowledge of, that the developers do have knowledge of
        - Common law negligence - Defendants' unreasonably dangerous designs and failure to exercise ordinary and reasonable care in its dealings with minor customers
        - There is a special relationship here by catering to children which created a special relationship. ACTIVELY solicited minors.

        - Negligence Per Se theory - violation of one or more state and/or federal laws prohibiting sexual abuse and/or online solicitation of minors and solicitation of unlawful acts, provision of mental health services without a license, and unlawful incitement of a minor to violent and/or unlawful acts
        - aiding and abetting liability - for design defect and failure to warn against Google. Claim against Google is that they provided material support. Claim is that Google knew of these issues and provided assistance (data, monetary support, underlying technology came from google, Google buys it back, LLM licensing)
        - COPPA Claim - repeatedly collected, used, or shared personal information about children under the age of 13. Did not obtain parental consent or notice to the parents. Need to NOTIFY and GET CONSENT

Mobley v Workday - Discrimination

- PARTIES
    - Mobley – Older Black gentleman with various disabilities (anxiety and depression) that applied to more than 100 jobs and got zero offers. Bachelors from Morehouse (HBCU).
    - Workday – Popular human capital management program
- ISSUE (Posture)
    - We are dealing with a Workday motion to dismiss. See above notes, i.e. *Iqbal, Twombley*.
    - Standard of review – interpreting the facts in the light most favorable to the non-moving party. Assume the facts as plead are true. Must suggest that the claim has at least a plausible chance of success. *Levitt v. Yelp! Inc*
- FACTS
    - Mobley applies to hundreds of positions and is universally declined. Allegation is that Workday is an agent of the employer (little more nuanced) and is bringing claims under the Age Discrimination in Enforcement Act, California Fair Housing and Employment, Civil Rights Act, and the Americans with Disability Act. The timing (immediately or in early morning hours) strongly implies an automated tool is making these decisions.
- DETERMINATIONS AT MOTION TO DISMISS
    - Court determines that Workday is an AGENT of the firms because it takes over "functions [that] are traditionally exercised by an employer." *Williams v. City of Montgomery* → this is agency liability
        - This is important. Without this, firms would only know about discrimination if the AI firm tells them about it
        - There is no artificial boundary between PEOPLE and AI AGENTS, because computers are making a lot of these decisions
    - Judge determines that Workday is not an Employment Agency because not curating positions and shopping them to people. It isn't finding people and matching them with jobs. Workday is a screening mechanism
    - Intentional discrimination: this is dismissed (failure to provide facts that this is intentional)
    - WRT California Employment and Housing:
        - Disposed of but permits continuation during discovery
        - Mobley does not allege behavior by any specific company
        - Mobley dos not argue that Workday knew its hiring decisions were discriminatory

Louis v SafeRent – Housing Discrimination

- FACTS

    - Woman (Louis) is rejected on a housing application due to an algorithmic grade she received. She appealed but the appeal wasn't heard by the renting company.
    - She provided evidence from prior landlords and the apartment housing firm simply declines
    - Louis is a Black woman with lower income and a housing voucher
    - DOJ submits a statement of interest saying that the algorithm could be held liable
        - SafeRent replies that they simply provide grades
- POSTURE
    - Motion to dismiss is denied. Case is then settled. Intuition suggests that as motion to dismiss is denied the firm settles rather than heading into discovery.
    - Settlement Terms – Cash, stipulation that the algorithm can't rely on Credit Score for people with housing vouchers, any revised algorithm must be approved by a third party endorsed by plaintiff
    - There is a little wobble in here due to proxy. Housing vouchers are ignored, and housing vouchers are meant to bridge the gap,
    - MY READ: THIS IS THE ISSUE WITH BLACK BOXING THINGS. IF REASONS WERE GIVEN AND THERE WAS AN APPEALS PROCESS THEY COULD PROBABLY AVOID THE LIABILITY

Young v NeoCortext – Deep fake consent

- PARTIES
    - Young – Actor whose likeness is included in deep-fake software that allows users to put their own likenesses into popular scenes from cinema / television (class action suit)
        - Young is a former Big Brother contestant
    - NeoCortext – Vendor of Reface application. Housed in Ukraine. Incorporated in Delaware
- FACTS
    - Reface deep fake tool is standard face swap technology. Free (watermarked) and paid versions. Issue is that consent was never given to use the actors' likenesses
    - Case is not about legality of the technology. It is about using likenesses for profit without obtaining consent. Case is brought under California Law (California Right of Publicity).
        - California law protects individuals from the unauthorized use of any of their attributes, including but not limited to, their names, signatures, photographs, images, likenesses, voices, or a substantially similar limitation of one or more of those attributes in the sale or advertisement of products, goods, merchandise, and services.
- SUBSEQUENT ACTIVITY
    - NeoCortext moved to dismiss under California anti-SLAPP statute, denied at district and appellate level
    - Both courts reject NeoCortext's claim of transformative use. Not an anti-SLAPP issue

Walters v Open AI - Defamation

- PARTIES
    - Walters – Just some guy from Georgia. Literally, just a guy
    - Open AI – I think we know what Open AI is at this point.
- FACTS
    - Journalist named Fred Reihl is investigating an ongoing suit between the Second Amendment Foundation, which is in the process of suing the Washington State AG (Ferguson) and the Washington State AAG (Studor)

    - When asked about the complaint, ChatGPT hallucinates and brings up Walters. Claims that Walters the treasurer and CFO of Second Amendment Foundation (he was not)
        - ChatGPT also claims that Walters was embezzling funds from the foundation
    - When Reihl asks for the complaint against Walters, ChatGPT hallucinates again and spits out a fake complaint with SAF suing Walters. It is making up FALSE allegations.
- POSTURE
    - This is a complaint, so it sort of ends there. The core claim is a defamation claim. There are four elements per *Paterson v. Little, Brown & Co. Mark v. Seattle Times.* Plaintiff bears burden of establishing a *prima facie* case on all four elements.
        - 1) falsity – statement must be demonstrably false. TRUTH is a complete defense
        - 2) an unprivileged communication – this is publication to a third party. The statement needs to be communicated to someone else.
        - 3) fault – hinges on public or private persons
            - Public Figures – "actual malice" is needed. *NY Times v Sullivan.*
            - Private Persons – negligence standard. Failure to exercise reasonable care
        - 4) damages – actual or presumed. Harmed reputation, economic losses. These are frequently presumed based on other elements.
    - FERRARO argues that this could have been a defective liabilities claim. See elements of Strict Liability claim above (Duty, Breach, Causation, Damages)
- NOTES
    - FERRARO comment. When you're lawyering, you are really trying to get into discovery for cases like this because that will be expensive and potentially embarrassing for defendant.

Woodruff v Oliver – Facial Recognition

- PARTIES
    - Portia Woodruff - 32-year-old pregnant woman (Black) who is arrested in front of her kids for robbery and vehicle theft. Held for multiple hours. Arraigned.
    - Oliver – LaShauntia Oliver is a detective with the Detroit PD. Assigned to the case
- FACTS
    - Tale as old as time. Guy picks up woman, they drink, have sex, he drops her off, and is robbed. Armed robber steals the man's car, wallet, and the phone. Police find the male robber but are looking for the woman.
    - Female accomplice is identified based on a surveillance photo of a person returning the phone to the same gas station. The woman who returns the phone looks a little like Woodruff (old picture from a booking for driving with a suspended license).
    - Police are working with a company called DataWorks Plus. That company feeds it through a database, finds a match with the prior booking photo. They do not use her driver's license photo, which they also have.
    - There is also a human in the loop. Human says they are a match. Police have the photo and do a six-person photo lineup with the victim who picks out the old booking photo.
    - Arrest occurs. Aside, this is a consistent issue, specifically for Black men
    - Woodruff charges are dropped because the perpetrator was not pregnant
    - What else could have been done (better system)
        - Downgrade odds with using older or low-quality photos
        - Cross reference with other public information
        - FERRARO brings up lighting structure of images. Others bring up "photo focal length"

        - ACLU argues that neither an FRT (Facial Recognition Technology) result, nor an eyewitness identification from a photo array based on an FRT result, should supply probable cause for an arrest. FRT results are fundamentally unreliable and display higher rates of false matches for people of color, women, and young adults

In the Matter of Rite-Aid Corporation – Facial Recognition

- PARTIES
    - RITE-AID – low rent drug store that is using FRT on customers
    - FTC – KHAAAAAAAN! Probably on the right side of history this time.
- FACTS
    - Rite-Aid is using facial recognition tech. False positives (shoplifting) has led to employees confronting people, being barred from shopping, employees without training calling the fuzz, all without notifying consumers and getting consent for facial recognition.
    - Another issue is that they are using low quality images which are dubiously sourced
        - Claim is a disproportionate effect on Black, Latino, and Asian persons
- POSTURE
    - Where does the rubber meet the road? FTC has conducted this investigation and is trying to get the judge to set a new order
        - This also implicates an earlier 2010 Data Security Issue (compounding problem)
    - FTC claims Rite Aid is engaging in unfair and deceptive trade practices.
- FTC Demands
    - Delete, and direct third parties to delete, any images or photos they collected because of Rite Aid's facial recognition system as well as any algorithms or other products that were developed using those images and photos;
    - Notify consumers when their biometric information is enrolled in a database used in connection with a biometric security or surveillance system and when Rite Aid takes some kind of action against them based on an output generated by such a system;
    - Investigate and respond in writing to consumer complaints about actions taken against consumers related to an automated biometric security or surveillance system;
    - Provide clear and conspicuous notice to consumers about the use of facial recognition or other biometric surveillance technology in its stores;
    - Delete any biometric information it collects within five years;
    - Implement a data security program to protect and secure personal information it collects, stores, and shares with its vendors;
    - Obtain independent third-party assessments of its information security program; and
    - Provide the Commission with an annual certification from its CEO documenting Rite Aid's adherence to the order's provisions.
    - For five years Rite Aid agrees not to use facial recognition, Rite Aid will get third party assessments.

People of California v Sol ECOM et al – Deepfake Pornography

- PARTIES
    - California – San Francisco District Attorney
    - Defendants – variety of deep fake companies. Some people. And a pile of John Does. They are in the US, New Mexico, Estonia. Slovakia.
- FACTS
    - Defendants are alleged to have created NCII pornography – nonconsenting persons face on another person's body (Fictitious Pornography)

    - Case is brought under California Consumer Protection statutes. Conduct happened in California (downloading). This essentially forbids any unfair, unlawful, or fraudulent business practice
    - Allegations regarding the models (LLMs)
        - Frequently Open Source – Allows users to download code and modify it to use it how you want. Open source is important to remove whatever safeguards do exist
        - Permits a user to deep nude another for free. You can sign in using a Google or Apple account. Website is initially free but additional use requires payment. Use terms require consent, but this is not safeguarded or enforced. The intent is clearly to circumnavigate consent (see anyone you know nude for free). Allegation is that this for adult and minor images
            - There is a mention of Google. FERRARO theorizes that this is probably a public shaming. Similar to pressuring credit card companies not to partner with revenge porn sites
- ISSUES
    - Development of AI models which are open source and can be used to create NCII
    - Operation of Websites that enable the nudification of people (women primarily but also minors)
        - Really comes down to the development and distribution
- CAUSES OF ACTION
    - VIOLATIONS OF BUSINESS AND PROFESSIONS CODE
        - California Civil Code (CCC) 1708.86(b)(1) prohibiting the creation and intentional disclosure of nonconsensual sexually explicit images, or aided and abetted violations
        - CCC 1708.85(a) prohibiting the intentional distribution of nonconsensual depictions of intimate body parts, or aided and abetted violations
        - CCC 647(j)(4) prohibiting the intentional distribution of nonconsensual depictions of intimate body parts of an identifiable person, or aided and abetted violations
        - 15 U.S.C. § 6851(b)(1) prohibiting the knowing or reckless disclosure in interstate commerce of intimate visual depictions of identifiable persons, or aided and abetted
        - Unfair business acts and practices (CCC 17200). Creating nudified images constitute unfair business practices because they offend established public policy, the harm they cause to consumers greatly outweighs any benefits associated with those practices, and they are immoral, unethical, oppressive, unscrupulous and/or substantially injurious to consumers.
    - Additional violations for such conduct for websites that create child images
- SIDE NOTE
    - Consider first amendment issues. Is there a first amendment defense? No way on the kids. For some of the rest there may bey a defense when moving towards things like nude fan art.
    - FERRARO notes that there is little distinction legally about public vs non-public figures but there are some legal attempts to distinguish between minors and adults

Random Cases

- *Mata v Avianca* – Airlines Lawsuit. Attorney builds the complaint using ChatGPT and it hallucinates cases.
- *In the Matter of Rytr LLC.* – FTC Action. AI is generating fake online reviews.
- *In the Matter of DONOTPAY, INC* - FTC complaint against RoboLawyer firm to assist consumers with civil issues (e.g. claiming rebates, cancelling subscriptions).

- *Doe v Github* pseudonymous software engineers filed a putative class action lawsuit against GitHub (Microsoft and OpenAI entities). Allegation is that defendants trained two generative AI tools—GitHub Copilot and OpenAI Codex—on copied copyrighted material and licensed code. Plaintiffs claim that these actions violate open source licenses and infringe IP rights. This litigation is considered the first putative class action case challenging the training and output of AI systems.
    - Case is dismissed. DMCA dismissal on appeal before the 9th Circuit
- *Anderson v Stability AI* - artists filed a putative class action lawsuit against Stability AI, Midjourney, and DeviantArt for copyright infringement over the unauthorized use of copyrighted images to train AI tools. The complaint describes AI image generators as "21st-century collage tools" that have used plaintiffs' artworks without consent or compensation to build the training sets that inform AI algorithms.
    - Currently in discovery
- *Getty Images v Stability AI* – Getty files against Stability, accusing it of infringing its copyrights by misusing millions of Getty photos to train its AI art-generation tool.
    - Currently in pre-trial awaiting renewed motion to dismiss