Addressing Emergent Regulations on Unbiased AI

Nick Maietta | VP of Legal Automation, Luminos.AI
nick@luminos.ai | Connect with me on LinkedIn:

Table of Contents

State of Play

Where We are Today

- Overall regulatory landscape is evolving, but shift towards subjective content standards

- Path dependency – did not get here in a vacuum

- Scalable capabilities of AI content creation can create fear of drowning out other competing perspectives

What are these Regulations About

- US: EO 14319
  - “Truth Seeking”
  - “Ideological Neutrality”
  - “[D]isclosure of the LLM's system prompt, specifications, evaluations, or other relevant documentation”
- EU: EU AI Act Articles 53 & 55
  - Measures to detect the unsuitability of data sources and methods to detect biases
  - Model evaluation using standardized protocols and tools reflecting the state of the art
- Others
  - South Korea: New, so still TBD, but AI ethics principles *as prescribed by Presidential Decree*
  - China: Respect social morality and ethics; adhere to the core values of socialism; not damaging the image of the country; not undermining national unity and social stability

Why this Matters

- Subjective standards often harder to comply with
  - May require more frequent review
  - Opens door to selective regulatory enforcement
- Moving towards to suitability on jurisdiction-by-jurisdiction basis?
  - Moving towards fragmentation?
  - Increased costs and complexity
- Public opinion can quickly become the story

Fundamental Challenges

1. How to effectively constrain the system’s output?

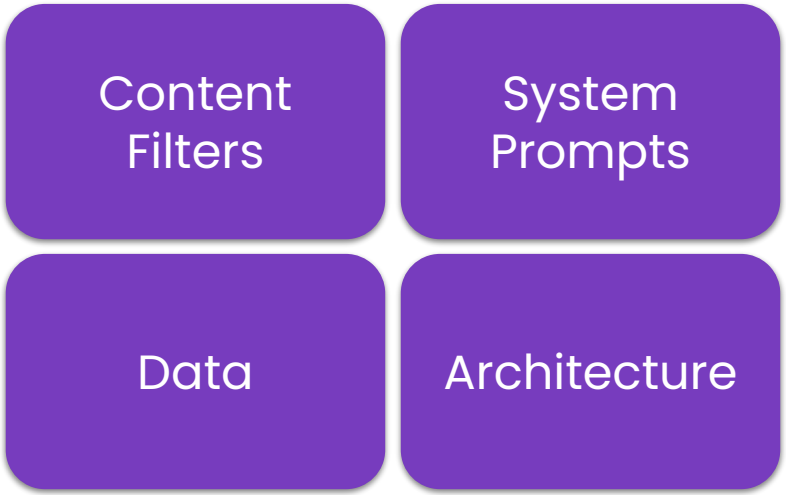
2. How do we know we are doing a good job?

3. How does this change for criteria with loose definitions?

Constraints

Taxonomy of Constraints

Not the whole universe, but to keep this practical, will stick to less esoteric approaches

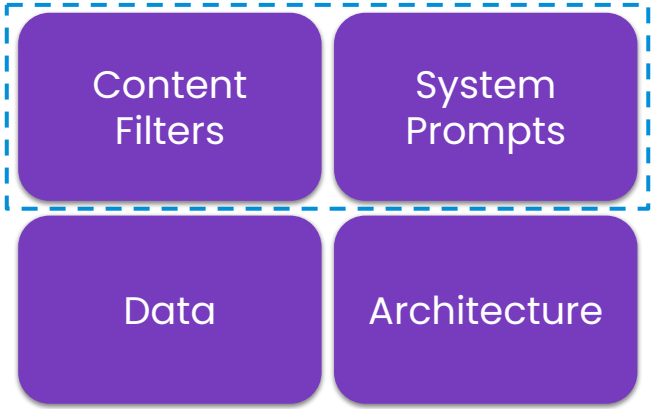


Constraints - Basic Approaches

Everyone starts here because of immediate impact and ease of implementation.

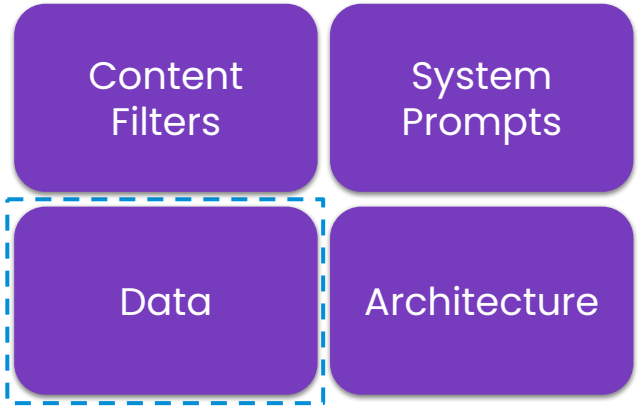
- **Content Filters**: Don't include "X"
  - Typically outputs, but can be inputs
  - Limits: Many (whack-a-mole, coded language, lack of context awareness, etc.)
- **System Prompts**: Tell system how to act
  - Limits: Many (unpredictable, context limitations, potential for bypass, etc.)

Could you defend if published?



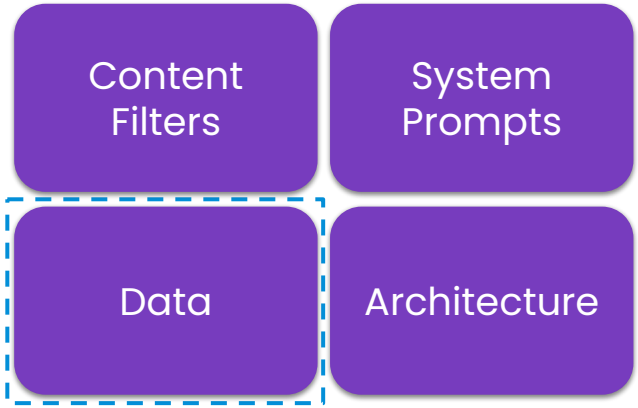
Constraints - Training Data

- May not be a lever you can pull . . .
- FM providers are focused on this
- Other jurisdictions have provided models for review
- Best practices still being developed, both from a practical standpoint, as well as from a legal perspective



Constraints - Fine Tuning

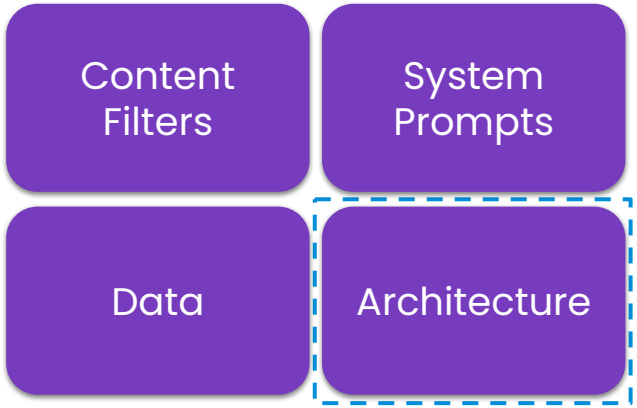
- Improves performance, but mixed (and potentially unpredictable) results on safety
- Performance: 👍
    - Improve domain expertise
    - Adhere to certain style/tone
    - Reduce certain hallucinations
- Safety: 🤦
    - Can lead to degradation of existing model alignment or guardrails
    - See this HAI Policy Brief for more
- Fine tune for performance, still need (at least) the same safety checks post-tuning



Constraints - Architecture

Includes both simple and complex approaches

- **System Parameters**
  - Temperature, top-k, top-p, etc.
  - Limits: Reduces utility (creativity, nuance, etc.)
- **Critical Review**: Various approaches
  - Reflection, pairwise comparison, etc.
  - Limits: Increase overhead (latency, inference), push towards norms; just problem-shifting to the reviewer?

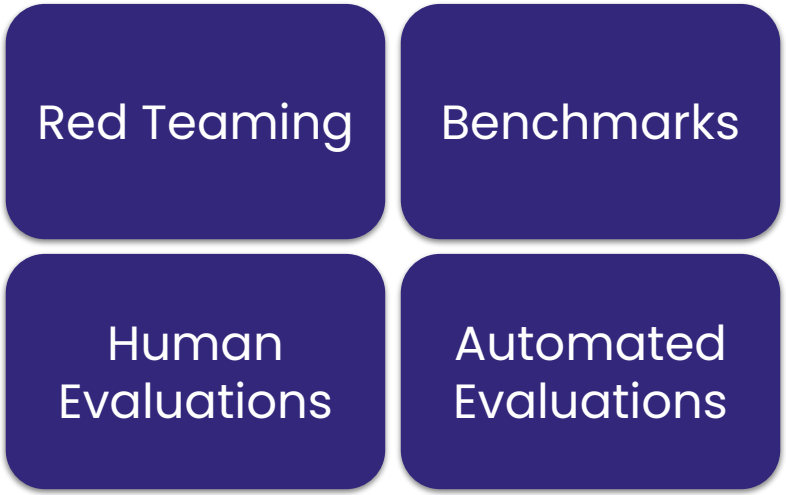


Who is the right stakeholder to manage critical review?

Measures

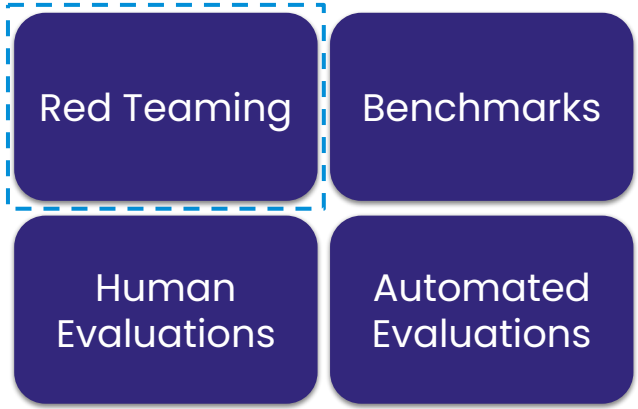
Taxonomy of Measures

Again, not the whole universe, but to keep this practical, will stick to less esoteric approaches



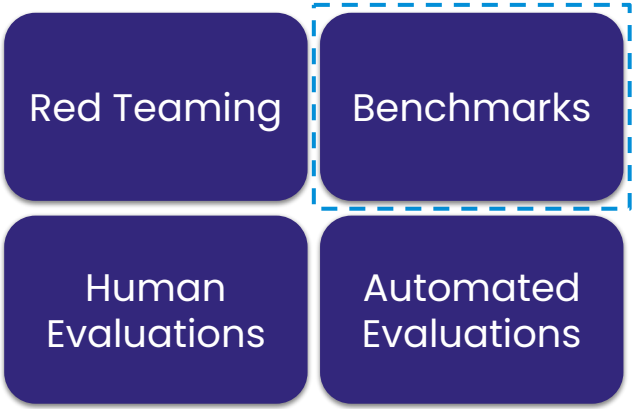
Measures - Red Teaming

- Can mean various things, but typically IDs vulnerabilities by designing prompts to avoid controls ("jailbreak")
  - Many parallels to the red teaming in the security space
  - Datasets of red team prompts available
- But breaking out of controls does not measure controls effectiveness when working
  - Two different problems
  - To carry forward the security analogy, "Who is doing blue/purple teaming"?



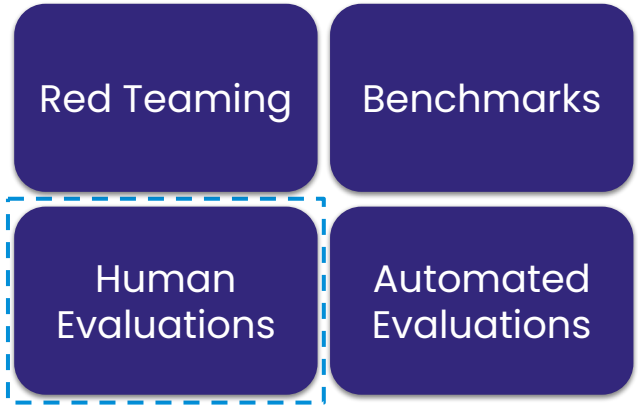
Measures Benchmarks

- Measuring performance of model against pre–defined dataset
- Can test different things:
  - Knowledge/Reasoning (e.g., MMLU, BIG-bench)
  - Safety (e.g., AdvBench, TruthfulQA)
- Various ways to implement:
  - Refusal to answer (e.g., OR-Bench)
  - Question answering (e.g., BBQ)
- But, see, Goodhart's Law
- Also, not tailored – why most useful for general purpose FMs



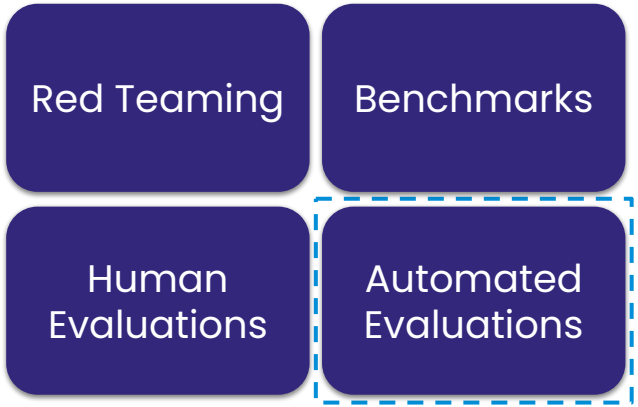
Measures - Human Evaluations

- Comparative human evaluations (i.e., which do humans prefer?)
- Most well known: LMArena (formerly Chatbot Arena)
- Who are the scorers? What are they evaluating on? What is their take on the subjective?
- If not tailored, then most useful for general purpose FMs



Measures - Automated Evaluations

- Use automated technology to scale evaluations
- Some of the same issues as human evaluations, but some different as well – how to bridge the best of bost worlds
- Technical hurdles to implementing with consistency
- If this seems interesting, then let's talk!



Addressing Subjectivity

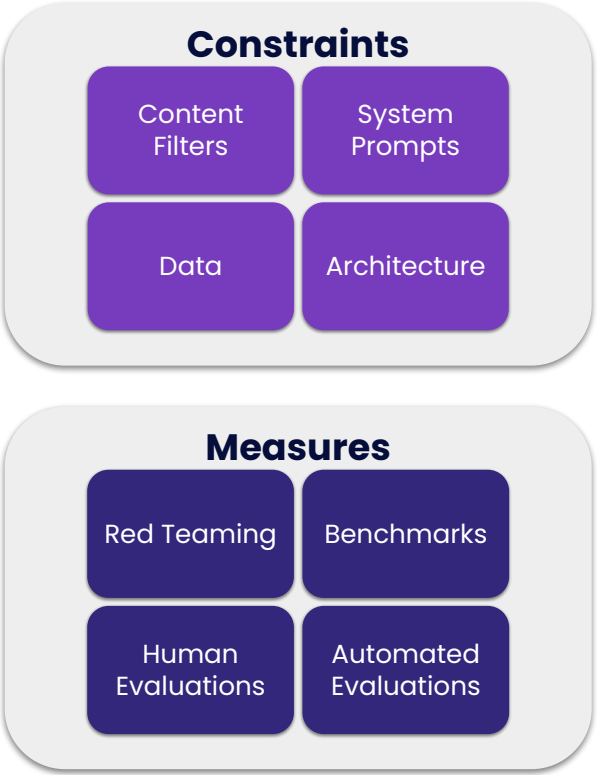
Role of Subjectivity

- Who is doing the measuring
  - Human – who is scoring; inherent preferences
  - LLM – also contain inherent preferences
  - Question: When would each preferable?

- What happens when the goal posts move?
  - Can be explicit (e.g., new guidance or regulatory action)
  - Or implicit (e.g., change in the zeitgeist)
  - Who is watching the goal posts?

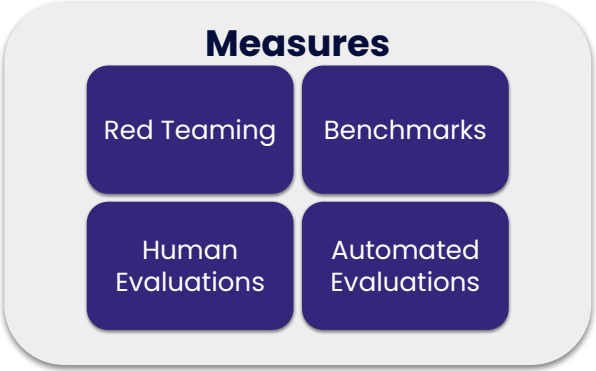
Relevant Terms from EO 14319

- Impartiality: How to measure?
  - Truthful
  - Historically Accurate
  - Objectivity
  - Neutral
  - Non-partisan
  - Acknowledge Uncertainty
- Actions: Limits on constraints?
  - *Do not manipulate responses* in favor of ideological dogmas
  - *Not intentionally encode partisan or ideological judgments* (unless prompted by or accessible to end user)



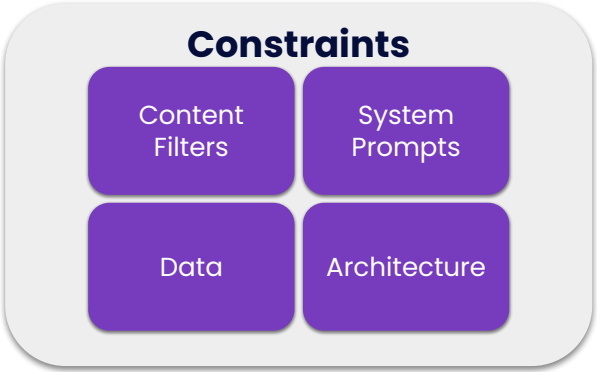
Impartiality - How to Measure

- Red Teaming: Appropriate expertise to identify impartiality?
  - How to align between client and red team?
- Benchmarks: Are “correct” answers actually impartial?
  - Build better benchmarks for impartiality?
- Human Evaluations: What biases are the human evaluators bringing?
  - Statistical techniques to better ensure representative evaluators? **Want to work on this – let’s connect!**
- Automated Evaluations: Is default FM behavior impartial to begin with?
  - If not, how to build evaluators without that bias?



Actions - Limits on Constraints

- Content Filters:
  - Are filters an encoding of ideological judgement?
  - Would need to be highly dynamic to keep up.
- System Prompts:
  - Relies on default tendencies of FM
  - Willing to expose? If not, is it “accessible” to end user?
- Data: Where can you source impartial data?
  - Some clearly better than others, but no bright lines.
  - Remove the worst and hope for the best?
- Architecture:
  - MoE to represent multiple sides of a position?
  - But of course, inference and latency . . .



Matching Evaluations to the Moment

- How to assess definitions?
  - Conservative vs aggressive readings?
  - In which direction?
  - When to re-evaluate?
- Really a variation on an old problem (regulatory interpretation) but what's different:
  - Technical nature of systems on which being implemented
  - Public-facing nature of many LLMs
- Who is best positioned to determine this for an organization:
  - Data scientists? Lawyers? PR?
  - Internal vs. External?
  - How to incorporate multiple stakeholders' input into these constraints and measures?
- A data scientist and a lawyer walk into a bar (link)

Thank You!

Nick Maietta | VP of Legal Automation, Luminos.AI
nick@luminos.ai | Connect with me on LinkedIn: