Formalized Policies - Efficient Regulation post-AGI

We regulate many human actions by evaluating them subjectively after the fact (e.g. litigation, adjudication); formalizing regulation could frontload the subjective evaluation allowing automated AI compliance verification.

Executive summary

If AI progress continues, responsible individuals (e.g. corporate officers, professional engineers, licenced attorneys) who review human proposals in high-stakes situations will be responsible for checking a deluge of AI-generated proposals against complex acceptance criteria (e.g. building codes, HR policies, legal contracts, tax codes, export controls, or security compliance requirements. The best solution is to formalize the policy to enable automated verification. Here a "formalized policy" has criteria expressed precisely and unambiguously so that they can be automatically checked (not subjectively interpreted). Atlas Computing is building AI-powered tools to test and facilitate formalizing policies to simplify compliance; we seek potential funders and early users.

The pending crisis of regulating AI actions

As AI systems become more useful, human review becomes more of a bottleneck. While users today review most AI outputs, more complex outputs will either be (1) processed slowly in an increasingly overwhelmed bottleneck, (2) reviewed in less depth, (3) delegated to untrustworthy AI models for review, or (4) reviewed in a robust automated way. We advocate for derisking the fourth path because review plays an important role (and therefore should not be abandoned), yet liability cannot be handed off to the AI model if/when review is.

For every AI system taking an important action, there will eventually be one person who checks inputs and is responsible for outcomes (e.g. corporate officer, professional engineer, or licensed attorney). Review could be handed off to AI but the liability cannot. But we could work to remove as much ambiguity as possible from the review process with practices from the field of formal methods, enabling safer automation.

Automate review by doing the work ahead of time

Laws, regulations, and other policies (like housing codes) could be considered descriptions of a nuanced boundary between scenarios that are allowed and those that aren't. Adjudication and review clarify where this boundary sits. However, identifying objective properties of allowed scenarios, enable automated validation

- e.g. imagine an architect demonstrates submits an online version of a building's plans including all metadata needed to automatically check compliance with the building code. This is not necessary but is clearly sufficient to automatically demonstrate compliance.

Policy formalization commonly guarantees safety-critical systems like aircraft autopilot and cryptography software. However, this process previously required formal logic experts to understand / translate the policies.

AI "experts" to formalize policies

As LLMs become increasingly capable, they could create formalizations of various policies. **We propose building an LLM-powered tool that generated a mechanized/computable policy from written policies and conversations with the experts who set and operationalize the subjective policy.** This could then be deployed incrementally to efficiently and safely automate checking of compliance with the policy.

- Formalization can be adopted incrementally; for instance, you could formalize a subset of the policies, and provide a "fast-lane" of review for proposals that prove that they comply with that subset.
- As AI systems grow in capabilities, they should should be able dramatically facilitate the formalization process, and be able to prove compliance against increasingly complex policies
- All these improvements should reduce regulatory compliance burdens for human actions as well.

Next steps

Email evan@atlascomputing.org if you're interested in using or funding a tool that could interactively formalize policies.