

Peer Review

Review of: "AI Ethics by Design: Implementing Customizable Guardrails for Responsible AI Development"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper argues that ethical guardrails for LLM systems should be configured by the people who use the system rather than fixed by the model provider, and it sketches an architecture for doing that. The building blocks are given in §4.2: three rule types, ordered by technical difficulty. Static rules are pattern matchers (the examples are regular expressions over email addresses, social security numbers, phone numbers, and other PII). Natural-language rules are sentences a non-technical user writes, with "never mention any content inappropriate for children below the age of 12" and "avoid conversation about religion" as the worked examples. Trained classifier rules are user-supplied or user-fine-tuned models, illustrated by a classifier that detects urgency in medical inquiries. Rules combine into policies (§4.3), evaluated sequentially with static first, natural-language second, classifiers last, and that order is itself user-editable. Policies combine into assistants (§4.4), which carry an input policy and an output policy, so the framework screens what the user sends as well as what the model returns. On violation, the configured action is redaction, blocking, a warning to the user, or escalation to a human.

Two further moves carry the argument. §5 positions the framework against Russell's three principles from *Human Compatible* [10], replacing "human preferences" with a user-curated, inspectable set of rules, on the ground that "human preference" is itself contested. §6 then treats conflicts between guardrails as a geometry problem: guardrails are said to have "ethical vectors as policy combination representations," and four cases are separated by whether the dot product is -1 or near it (§6.2 and §6.4 both use -0.9), and by whether the opposition is permanent or conditional. Five resolution strategies follow in §6.5: weighted averaging, strict precedence, a hybrid that falls back to precedence when mutual negation is detected,

contextual triggering, and handing the decision to a person. The abstract also states that the work compares the proposed framework "to the existing state-of-the-art guardrails"; that comparison is §3, a prose discussion of NeMo Guardrails [26], LlamaGuard [9], and Guardrails AI [28].

Credit where it is due. The insistence that policies apply to user input and not only to model output is the sharpest idea in the paper, and §4.4's framing of it is clean. The three-tier rule typology in §4.2 is genuinely useful as a design vocabulary because it separates what a regex can do from what needs a judgment call, and it is honest about which tier a non-specialist will actually reach for. §2 scopes out machine consciousness and metaethical prescriptivity explicitly instead of drifting into them, which many papers in this area fail to do. The writing is clear, and the philosophical citations are accurate where I could check them.

2. Whether the conclusion holds on the data

2.1 There is no data. I want to state this plainly rather than bury it because it governs everything else. The paper contains no experiment, no table, no benchmark, no user study, no measurement of any kind. The only figure is Figure 1, a block diagram of the architecture. So the ordinary questions in this section—sample size, controls, baselines, statistics—do not apply, and the question becomes which claims in the paper require evidence that is not present. Several do.

2.2 The abstract's comparison claim is not met. "Comparing the proposed framework to the existing state-of-the-art guardrails" sets up an expectation that §3 does not satisfy. §3 describes three systems and then asserts their shortcomings, but there is no comparison along fixed axes, no feature matrix, and above all no case where the same input is run through NeMo, LlamaGuard, Guardrails AI, and the proposed prototype so a reader can see the difference. A qualitative comparison is acceptable in a position paper. Calling it a comparison to the state of the art in the abstract, when the proposed system is never placed beside the others on any common input, is not.

2.3 The critique of competitors rests almost entirely on one secondary source. I traced the citations in §3. The account of NeMo's embedding plus K-nearest-neighbor matching, the reliability objection to LlamaGuard ("the classification results depend on the model's 'understanding' of the categories and its predictive accuracy"), and the text-only limitation of Guardrails AI are all attributed to [19], the Dong et al. survey. That is three competitor assessments from one survey, with no primary verification. The paragraph beginning "While NeMo Guardrails, LlamaGuard, and Guardrails AI offer valuable safeguards" and the three paragraphs after it carry no citations at all, yet they make checkable empirical claims: that

NeMo "involves increased token usage for their prompt engineering-focused solution, which leads to higher operating costs"; that the alternatives "lack the flexibility to quickly adapt to new data privacy regulations"; that open-source solutions are "better geared for assisting individual developers or small companies." The token-usage claim in particular is easy to measure and would take an afternoon. Measuring it would strengthen the paper considerably, since cost is the one axis on which the proposal could show an advantage without a full safety evaluation.

2.4 The vector formalism in §6 is never defined. This is the objection I consider most serious. §6.1 speaks of "their ethical vectors as policy combination representations" having "a dot product of -1," and §6.2 and §6.4 use -0.9. Nowhere in the paper is it said what space these vectors live in, how many dimensions they have, how a policy written as a list of natural-language sentences is mapped into one, whether the vectors are normalized (the dot-product thresholds only make sense if they are), or who computes them. §6.1 further claims such opposition "could be identified through static analysis," but no analysis is specified. Without a definition, the four-case taxonomy of §6.1 to §6.4 is a set of labels for situations nobody can detect, and the weighted-averaging strategy of §6.5 has nothing to average. The privacy-versus-transparency example given in §6.1 also does not obviously have a dot product of -1 under any embedding I can imagine: those two values conflict in some contexts and are orthogonal in most.

2.5 §6 does not operate on the object built in §4. §4.3 defines a policy as an ordered chain of rules, evaluated sequentially with a deterministic pass or fail at each step. §6 treats guardrails as points in a continuous space that can be averaged. These are two different machines, and the paper never reconciles them. If a static PII regex fires and a natural-language rule says "continue the conversation naturally," what does weighted averaging mean? Redaction and blocking (§4.4) are discrete actions. Averaging them is not defined. The conflict machinery would need to be rewritten in terms of the rule chain, or the rule chain rewritten in terms of vectors.

2.6 Two claims about conflict frequency contradict each other. At the end of §3, the paper asserts that "very rarely one will find practical examples of contradictions between the end user's ethical values when applicable to AI guardrails," and this assertion is what licenses the whole approach: if user-level value conflicts were common, pushing the ethics down to the user would just relocate the problem. No evidence is offered for it. Then §6 spends four subsections and five resolution strategies on exactly those conflicts, and the closing paragraph of §6.5 calls Cases 1 and 3 "significant challenges." Both positions

cannot stand. My own guess, stated as a guess, is that the §6 view is the right one and the §3 sentence should go.

2.7 The circularity the paper raises is not escaped. §4.2 states the worry well: "in order not to fall in the same trap of using a LLM to mitigate responses from an LLM." Two escapes are offered. The first is keyword lists, sentiment analysis, and lexicon-based methods. The second is self-hosting Llama or Mixtral. The second is not an escape from the trap as stated; it only changes who runs the second model; the data-custody benefit is real, but it is a different benefit. The first escape is asserted without support, and it is the load-bearing claim of the paper. A keyword and lexicon pipeline enforcing "never mention any content inappropriate for children below the age of 12" is precisely the kind of rule where lexical methods are weakest, since the offending content is usually a matter of implication rather than vocabulary. One small experiment would settle this: take twenty natural-language rules of the kind §4.2 lists, write a few hundred test utterances, and report how often the lexicon route agrees with human judgment.

2.8 Error costs are never discussed. For a safety mechanism, this is a gap. No rule type in §4.2 comes with any treatment of false positives or false negatives, and the medical-urgency classifier in §4.2 is the example where it matters most: a missed urgent inquiry and a wrongly escalated routine one have very different costs, and a user configuring their own classifier from "a few synthetic few-shot examples" (§4.2) has no way to know where the threshold sits. The blocking action in §4.4 can cut off a conversation entirely, so a false positive is not free either.

2.9 The threat model raised in §1 is dropped. §1 introduces prompt injection as a motivating problem, citing [7], a paper by two of the present authors. The framework in §4 never returns to it. Natural-language rules are themselves instructions in natural language, so the obvious question is whether an injected instruction in the user turn can suppress or reorder them, particularly given that §4.3 lets the evaluation order be reconfigured. The paper should either address this or say that adversarial resistance is out of scope, which would be a defensible choice stated openly.

2.10 The three principles in §5 need one more pass. Principle 2 says "the machine and user are initially uncertain about what those preferences are," which is a real improvement on Russell's formulation, since it admits the user has not introspected their own values. But principle 3 then makes the curated rule set the ultimate source of information about the user's preferences, which drops Russell's point that behavior, not stated preference, is the evidence. That is a substantive departure, and it deserves a

sentence defending it rather than passing unremarked, because a rule set is exactly the kind of stated preference that comes apart from what people do.

3. Reproducibility

The paper is a design paper, so the relevant standard is not seeds and checkpoints but whether a competent reader could build the described system. I do not think they could, and the specific gaps are these.

No artifact. There is no code repository, no configuration schema, no pseudocode, no formal grammar for rules, and no machine-readable example of a policy. The single pointer is footnote 15, "Preamble AI Trust Platform, app.preamble.com (2024)", a hosted product behind a sign-up, with no version identifier and no archived snapshot. §4.5 calls this "the prototype in action", but a live commercial endpoint is not a reproducible artifact: whatever a reader sees there in 2026 need not be what the authors saw in 2024.

No model versions or parameters anywhere. This is the usual failing of LLM papers, and it holds here in full. The paper names Llama and Mixtral in §4.2 as self-hosting options with no size, no version, and no snapshot date. No LLM is named for any part of the described pipeline. No decoding parameters, no temperature, no context limits, no prompts, no seeds, no compute figures appear at any point. Since the paper criticizes NeMo for token usage, the absence of its own token accounting is doubly noticeable.

Classifier rules are underspecified. §4.2 says users can upload their own datasets, use public ones for fine-tuning, or have an LLM generate synthetic few-shot examples. No base model family, no minimum dataset size, no training procedure, no decision thresholds, no validation protocol. The claim that classifiers "when provided with enough data, can offer a more nuanced and precise level of control" has no quantity attached to "enough".

Undefined terms in the architecture. §4.4 defines an assistant as a combination of input and output policies "along with preferred system prompts and action items". Action items are not defined here or anywhere else in the paper. Similarly, §4.3 says the default rule order "can be altered by the user to create a hierarchical chain", but the semantics of a chain are not given: whether evaluation short-circuits on first violation, whether later rules can override an earlier redaction, and what happens when a redaction rule and a blocking rule both fire on the same span.

Figure 1. As rendered on Qeios, only the caption of Figure 1 reaches the text I was given; the diagram itself is not legible in the version I reviewed, so I cannot say whether it resolves any of the ambiguities above. The caption contains a typographical error, "static ules".

Reference integrity. I checked the identifiers in the reference list and found collisions. [19] (Dong et al., "Building Guardrails for Large Language Models") and [20] (Wang and Singh, "Adding Guardrails to Advanced Chatbots") are both given as arXiv:2306.07500. [23] (Lei et al., "New Challenges in Reinforcement Learning: A Survey of Security and Privacy") and [26] (Rebedea et al., "NeMo Guardrails") are both given as arXiv:2310.10501. Two different papers cannot share a preprint identifier, so at least one entry in each pair points at the wrong document, and a reader following [26] to check the NeMo claims will land somewhere else. Separately, [9] and [27] are the same work: [9] is the Llama Guard announcement page with "Meta.com" in the author field, [27] is Inan et al., arXiv:2312.06674. The body text uses both numbers for the same system within §3. Finally, the EU material quoted in §1, "'trustworthy AI' should be lawful, ethical, and technically robust", plus the list of seven requirements, carries no citation at all, and it is a direct quotation.

Declarations. In the version rendered on Qeios, I can read only the Funding statement under Statements and Declarations ("The authors received no funding for this work"). The corresponding address is jeremy@preamble.com, and footnote 15 points to a Preamble product, so the paper both proposes a framework and points at a commercial implementation of it by the authors' organization. If a competing-interest statement exists behind the collapsed Declarations control, I could not read it, and it should be visible in the article text.

4. *What specifically to fix*

1. **§6.1 to §6.4, the vector formalism.** Define the space. Say what the components of an "ethical vector" are, how a policy is mapped to one, whether vectors are unit-normalized, and by what procedure the dot product is computed. If the intent was a metaphor rather than a computation, say so and drop the numbers -1 and -0.9, because as printed they read as quantities. If the intent was a computation, give the embedding and one worked example with actual components for the privacy-versus-transparency pair of §6.1.
2. **Reconcile §6 with §4.3.** The sequential rule chain and the averaging of vectors are two incompatible execution models. Pick one and rewrite the other in its terms. At minimum, state what "weighted averaging" (§6.5, first strategy) yields when the competing actions are redaction and blocking, which are discrete.
3. **Abstract, sentence 3.** Either downgrade "comparing the proposed framework to the existing state-of-the-art guardrails" to a claim about a conceptual discussion, or add the comparison: a table with

NeMo, LlamaGuard, Guardrails AI, and this framework as columns and fixed axes as rows (input-side coverage, multimodality, customization entry point, technical skill required, token overhead, licensing).

4. **§3, the four uncited paragraphs starting at "While NeMo Guardrails, LlamaGuard, and Guardrails AI offer valuable safeguards".** Every empirical claim there needs a source or a measurement. The NeMo token-cost claim is the one I would measure rather than cite: run a fixed set of prompts through NeMo and through the proposed chain, report input and output tokens per turn.
5. **Delete or defend the sentence in §3 beginning "very rarely one will find practical examples of contradictions between the end user's ethical values".** It contradicts the premise of §6.
6. **§4.2, the non-LLM enforcement route.** The claim that keyword lists, sentiment analysis, and lexicon-based methods can enforce natural-language rules needs a small evaluation. Twenty rules of the type already listed in the section, a few hundred labeled utterances, precision and recall per rule type. This is the single addition that would most change my assessment of the paper.
7. **§4.2 and §4.4: add a paragraph on error costs.** The medical-urgency classifier example is where to anchor it. What is the cost of a false negative, what is the cost of a blocked conversation, and what guidance does a non-technical user get for setting the threshold?
8. **Prompt injection.** §1 raises it via [7]; §4 should return to it. Whether natural-language rules survive an injected instruction in the user turn is the first question a security reader will ask. An explicit out-of-scope statement would be acceptable.
9. **References [19] and [20] both carry arXiv:2306.07500; [23] and [26] both carry arXiv:2310.10501.** Fix both collisions. Merge [9] and [27], which are the same Llama Guard work, and repair the author field of [9], currently "Meta.com".
10. **§1, the EU quotation.** Add the citation for "'trustworthy AI' should be lawful, ethical, and technically robust" and for the seven requirements that follow.
11. **§4.4.** Define "action items". Specify chain semantics: short-circuit or not, override rules, and the outcome when redaction and blocking fire on the same content.
12. **Footnote 15 and §4.5.** A live URL is not an artifact. Add a version or access date for the platform, and if any part of the rule engine can be released, release it; a schema and a handful of example policies in a repository would be enough to make §4 implementable.
13. **§5, principle 3.** Add a short defense of replacing Russell's "human behavior" with a user-curated rule set as the ultimate source of preference information, since stated rules and revealed behavior routinely diverge.

14. **Small textual errors.** §6.3 reads "Variants I, III, and III from Case 1 apply here as well"; the second "III" should presumably be "II". §6.5, third strategy, reads "(e.g., in Case 1/I, 1/III, 3/I, or 3/I)" with 3/I listed twice. The caption of Figure 1 has "static ules". §5 has "A practical ethical AI agents" and "a range of resolution mechanism" appears in §7.

5. Recommendation

Major revisions.

I think the core position is worth publishing. Moving ethical configuration to the user, screening input as well as output, and separating rules by the technical skill they demand are all sensible, and the paper states them more clearly than most work in this area. The philosophical framing in §2 and §5 is careful, and the authors are candid about what they are not trying to settle.

What stops me short of a lighter verdict is that the paper's two central mechanisms are not specified well enough to be built or checked. The conflict machinery in §6 uses a quantitative vocabulary with no quantities behind it, and it does not act on the architecture defined in §4. The claim in §4.2 that natural-language rules can be enforced without a second language model is the hinge of the whole design and carries no support. Neither of these needs a large study to repair: a definition and one worked example for the first, a small labelled evaluation for the second. The reference collisions and the uncited EU quotation are mechanical fixes.

If the vector formalism is defined or withdrawn, if §6 is expressed in terms of the rule chain, and if the enforcement claim in §4.2 gets even a modest empirical backing, I would read a revised version favorably.

Declarations

Potential competing interests: No potential competing interests to declare.