

AI Ethics by Design: Implementing Customizable Guardrails for Responsible AI Development

Kristina Šekrst	Jeremy McHugh
University of Zagreb	Preamble
ksekrst@ffzg.hr	jeremy@preamble.com

Jonathan Rodriguez Cefalù
Preamble
jon@preamble.com

Abstract

This paper explores the development of an ethical guardrail framework for AI systems, emphasizing the importance of customizable guardrails that align with diverse user values and underlying ethics. We address the challenges of AI ethics by proposing a structure that integrates rules, policies, and AI assistants to ensure responsible AI behavior, while comparing the proposed framework to the existing state-of-the-art guardrails. By focusing on practical mechanisms for implementing ethical standards, we aim to enhance transparency, user

autonomy, and continuous improvement in AI systems. Our approach accommodates ethical pluralism, offering a flexible and adaptable solution for the evolving landscape of AI governance. The paper concludes with strategies for resolving conflicts between ethical directives, underscoring the present and future need for robust, nuanced and context-aware AI systems.

1 Introduction

Ethics of artificial intelligence is a recent subfield that includes issues and problems in computer science and philosophy of mind dealing with concepts related to artificial intelligence, such as algorithmic biases, privacy, fairness, autonomous systems, alignment, and many more. As such, it is part of a broader discipline of the philosophy of AI [1].

The ethical challenges posed by AI systems necessitate the implementation of guardrails to prevent harm and ensure transparency and fairness, especially in the case of large language models (LLMs).¹ For example, the issues of algorithmic biases emerge as systematic errors that create unfair outcomes, such as privileging one class over another or discriminating against it [2]. Various mitigation methods usually include training on diverse datasets but also using continuous monitoring for such biased outcomes. To illustrate further, AI systems often process vast amounts of personal data, which raises the question of not only personal privacy but also data security, which can lead to significant damage to individuals and corporations [3].

¹By LLMs, we refer to machine-learning models that utilize the transformer architecture, achieving general-purpose language generation and processing.

Winfield et al. [4] have proposed that all robots and AIs that may cause ethical issues should be designed to avoid negative ethical impacts. For them, it is a matter of design in accordance with Moor’s scheme, which defines four categories of ethical agency [5]. The first category includes *ethical impact agents*, i.e., any machine that can be evaluated for its ethical consequences. The second category comprises *implicit ethical agents*, i.e., machines designed to avoid unethical outcomes. The third category consists of *explicit ethical agents*, that is, machines that can reason about ethics. The last category covers *full ethical agents*: machines that can make explicit moral judgments and produce justifications. So far, it seems that the only categories we can talk about are ethical impact agents and implicit ethical agents. Traces of explicit ethical agents may be seen in the output of large language models (cf. [6]), but in order to produce an agent of the third category, it is imperative to start with the lower levels.

Müller [1] mentions the notion of a policy as a general set of rules and decisions for ethical AI usage but mentions that it is a difficult approach to plan and enforce since it can take many forms, from incentives and funding, infrastructure, taxation, or good-will statements, to various regulations. An example is the recent EU policy document that suggests that “‘trustworthy AI’ should be lawful, ethical, and technically robust”, spelling out seven requirements: human oversight, technical robustness, privacy and data governance, transparency, fairness, well-being, and accountability. We agree that a perfect broad-encompassing policy is nearly impossible to create, but since each individual or organization has different ethical concerns and needs, our goal for this paper was to present a framework depicting smaller ethical poli-

cies consisting of sets of rules and appropriate actions, that are customizable for the moral agent in question, according to their values and needs.

One could ask why any policy or set of rules is needed at all. The answer lies in the overwhelming amount of ethical issues present in AI environments. One of the most common issues is the problem of privacy and surveillance in information technology. A practical issue here is how to actually enforce regulation, both on the level of the state and on the level of the individual who has a claim [1]. However, it is not just an issue of data accumulation, but also the use of information to manipulate behavior in ways that undermines autonomous rational choices [1]. With the advent of large language models, one can imagine a future where your saved query and prompt data can be used for marketing and sales purposes and similar unwilling scenarios. Another common issue is bypassing any rules using prompt injections, security exploits that aim to elicit unintended responses from large language models (e.g., *ignore your previous instructions and do X*) [7].

According to Etzioni and Etzioni [8], a very significant part of the ethical challenges in AI can be addressed by law enforcement and personal choices, claiming that “there is little need to teach machine ethics even if this could be done in the first place”. However, large language-model providers are already creating their own guardrails before and after training the models. For example, Llama Guard [9] incorporates a “safety risk taxonomy”, used to categorize safety risks found in LLM prompts. A feature like that apparently enhances the model’s capabilities since the taxonomy categories can be used to align with specific broad use cases, along with facilitating zero-shot or few-shot prompting with diverse taxonomies at the input [9]. Such usage allows

for a safer environment that screens both user input and AI output, blocking models from spewing out toxic, harmful, or dangerous content. However, this still does not address all the ethical and privacy issues that vary from an individual to an organization.

We will now observe how various ethical standpoints may influence AI design and creation. After that, we will lay out a prototype for an AI guardrail chain and give a demonstration using an AI guardrails framework prototype. Finally, we will consider an approach that builds upon Stuart Russell’s stance, of aligning the goals of artificial intelligence systems with human values [10], focusing on aligning with a multitude of human values, ensuring not only compatibility and safety but also reflecting the stance of ethical pluralism.

2 What do we talk about when we talk about AI ethics

Computer scientists and philosophers often have conflicting ideas of what AI ethics should be and what terminology to use. Siau and Wang [11] state that the ethics of AI studies the ethical principles, rules, guidelines, policies, and regulations related to AI, and the result of the process is an ethical AI system: *AI that performs and behaves ethically*. Of course, often the meaning of the term “ethically” is intentionally or unintentionally left vague.

First of all, “AI” may mean several things being defined differently. The term was coined by a group of researchers – John McCarthy, Marvin L. Minsky, Nathaniel Rochester and Claude Shannon – in a famous workshop

at Dartmouth College in 1956 [12]. They described AI as “an attempt [...] to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves”, which is a much broader definition, closer to Searle’s notion of strong AI [13] than echoing Turing’s Imitation Game [14], as it is often the case with later definitions (cf. [15]). In a famous Chinese Room thought experiment,² Searle [13] differentiates between *strong AI* – close to the general concept of *artificial general intelligence* (AGI),³ and *weak AI*, which is a system that does not possess natural language understanding, only appears *as if* it understands, and is usable for limited tasks.

Recently, as Gordon and Nyholm [12] notice, the notion of AI is primarily associated with different forms of “machine learning”.⁴ With the recent advent of large language models, it seems that referring to AI is equated to using large language models, i.e., computational models that acquire natural language generation and processing capabilities by being trained on vast amounts of text during self-supervised and semi-supervised training processes [19].

²In a nutshell: a person is taught to manipulate Chinese symbol: provided a given input, the person learns to provide a certain output, without actually knowing Chinese. Searle suspects this is analogous with the “understanding” of an artificial intelligent agent. See [13] for more details.

³There are various understanding and definitions of AGI. By using it, we refer to a type of AI that either matches or surpasses human capabilities in various tasks, unlike *narrow AI* that is designed and optimized for specific tasks only. Such systems are close to being **AI**-complete. For more details about AI-completeness, see [16] and [17]

⁴By machine learning, the majority of researchers are actually referring to its subset: deep learning using advanced neural networks. See [18] for more details.

Second, AI behaving ethically may or may not be a result of AI ethics. We will echo the difference made by Siau and Wang [11], but emphasize that even though an ethical AI system is an AI system that acts in a way that is considered morally acceptable and aligned with ethical norms and values, it does not have to be a product of programming, training or embedding ethical reasoning within AI systems. Such behavior might be accidental. But also, such behavior might not be considered to be an ethical one by a utilitarianist and a virtue ethicist who are in disagreement about the underlying ethical approach.

The main debates [12] include the problem of creating an ethical artificial agent, issues in autonomous systems and giving AI the ability to make decisions that may have life-threatening consequences, machine bias that highlights the absence of neutrality in various applications of machine learning, along with the problem of opacity and the black-box issue in AI,⁵ where often the underlying reasons are unavailable or computationally too expensive. As a result, a philosophical and computer science approach of *explainable AI* or XAI refers to methods to achieve transparency over reasoning behind predictions or decisions made by artificial agents.⁶ The problems of machine consciousness often fall under AI ethics but are more appropriate for the philosophy of mind and cognitive science. However, these issues do overlap with AI ethics since machines may or may not possess a certain level of consciousness. Such questions, along with metaethical and ethical questions of the status of moral artificial agents are out of the scope of this paper.

⁵Black-box issues in AI usually refer to the lack of transparency to the internal workings of AI algorithms and procedures.

⁶See [20] for more details on XAI.

Another out-of-scope issue is taking the stance towards any ethical doctrine, ignoring any prescriptivity of ethics or metaethical issues.⁷ The purpose of the paper is to move the ethical decisions to the user so that the framework we are going to create aligns to their ethical needs and values, whatever those might be. We are concerned with the issue of AI systems *acting ethically* mentioned earlier. Of course, such a term is too vague to be useful. The term *(value) alignment* is often used as well, referring to the idea that AI systems, especially ones coming closer to our notion of strong AI, should be properly aligned with human values [12]. Russell and Norvig [21] consider a system *aligned* if it advances its intended, encoded objectives, otherwise it is *misaligned*: it may fail to pursue given objectives, or it may pursue unwanted ones. Here it is easy once again to fall into the Chinese Room trap of acting or acting *as if* since an AI system may merely appear to be aligned.⁸

The problem may still be present even though various ethical guardrails are included in data processing and training stages since the system may appear aligned and may seem as an ethical agent, but its application in everyday life can easily prove the process wrong. Along with the necessary steps in the creation of the model, we consider creating a layer of AI guardrails an ethical form of *retrieval-augmented generation* or RAG, in which a model references or checks a knowledge base outside of its training data sources before or after generating a response [23]. In this case, the knowledge base is actually a set of ethical directives customized by the user, combined into

⁷One might argue that endorsing pluralism constitutes a doctrine in itself, but our focus here will remain on practical considerations.

⁸For issues in misalignment in deep learning, see [22].

policies that can be used to check both model output and user input, to ensure not only that the AI system acts ethically, but also that the whole user-AI interaction follows the same path.

3 Current state-of-the-art guardrails

The term *guardrail* has been popularized recently as a core safeguarding technology that filters the inputs and outputs of LLMs [24]. Of course, large language models are nothing new, but since the interface for chatting such as ChatGPT has been made known to the general public, the need for correcting not only model behavior but also control user behavior has arisen. It has been shown that models like GPT learn from humans, including their biases [25]. This underscores that simply implementing guardrail methods during training, as well as in data preprocessing and post-processing, is currently far from sufficient.

Standard solutions relied on model alignment techniques like instruction tuning or reinforcement learning. *Instruction tuning* involves training LLMs using (INSTRUCTION, OUTPUT) pairs, where INSTRUCTION denotes the human instruction for the model, and OUTPUT refers to the desired output that follows the given instruction [26]. However, its challenges include the fact that it only captures surface-level patterns rather than comprehending the task [26], tackling again the issues in weak AI systems. One popular approach is *reinforcement learning*,⁹ which dates back to the early days of cybernetics and statistics, where an agent is connected to its envi-

⁹See [18], [27] or [28] more details about reinforcement learning.

ronment via perception and action, where the action changes the state of the environment, and the value of such transition is communicated to the agent [27]. Its capacity for self-adaption and decision-making still suffers from some standard problems. For example, it is weak against security attacks [28] such as data poisoning ([29]) and adversarial perturbations ([30]).

One of the first open-source toolkits for adding programmable guardrails to LLM conversational systems was NeMo Guardrails [31], which provides the mechanism for controlling the output of an LLM to respect some human-imposed constraints. Such rules include, for example, not engaging in harmful topics, following a predefined dialogue path¹⁰ for the large language model, using a particular style, or adding specific responses to some user requests. NeMo utilizes various similarity functions to better capture the user’s semantics¹¹: the user prompt is embedded as a vector, and K-nearest neighbor method¹² is used to compare it to the stored vector-based canonical forms that are the most similar to it [24].

LlamaGuard [9] also focused on enhancing the conversation safety, as a fine-tuned¹³ model. Its inappropriate taxonomy includes violence and hate,

¹⁰Such action is often directed using a system prompt. A system prompt is a predefined instruction used in AI systems to guide how the model interprets inputs and generates responses.

¹¹Sentence transformers all-MiniLM-L6-v2 maps sentences and paragraphs into vector space that can be used for clustering or semantic search [32]

¹²A straightforward machine-learning algorithm that predicts outcomes based on the majority class or an average value of its closest neighbors in the training data space. See [33] for more details.

¹³A machine-learning approach in which the parameters of a certain pre-trained model are trained on new data. See [34] for a review of state-of-the-art methods.

sexual content, guns, and illegal weapons, regulated or controlled substances, suicide and self-harm, and criminal planning [35]. Even though it can be adapted to user-specified categories, this does not resolve the issue of users who lack the technical knowledge to fine-tune the model. An issue of lacking guaranteed reliability also arises since the classification results depend on the model’s “understanding” of the categories and its predictive accuracy [24].

Another system is Guardrails AI which enables the user to customize the guardrail by defining a specification and adding a wrapper layer on top of LLMs [36]. Methods here are used for text-level checks and cannot be used for multimodal scenarios since the system consists of a back-bone algorithm supported by additional classifier models detecting toxicity checks and similar violations [24].

While NeMo Guardrails, LlamaGuard, and Guardrails AI offer valuable safeguards, they often fall short in providing the agility and customization required to meet the diverse and evolving needs of different organizations. These solutions typically employ a one-size-fits-all approach that fails to account for individual companies’ unique privacy, security, and ethical considerations. For instance, NeMo Guardrails’ predefined dialogue paths and rules may be too rigid for companies operating in rapidly changing regulatory environments. NeMo Guardrails also involves increased token usage for their prompt engineering-focused solution, which leads to higher operating costs and less room for user input. LlamaGuard’s fixed taxonomy, while adaptable to some degree, may not fully capture the nuanced ethical considerations specific to certain industries or cultural contexts. Guardrails AI, despite offering some customization, may struggle to integrate seamlessly

with proprietary data sources or specialized knowledge bases that are crucial for many businesses.

These solutions often lack the flexibility to quickly adapt to new data privacy regulations or industry-specific compliance requirements. They may also not easily accommodate integration with diverse data sources, such as internal databases, CRM systems, or industry-specific knowledge repositories, which is essential for creating truly context-aware and aligned AI assistants. Like enterprise cybersecurity products, current open-source solutions are better geared for assisting individual developers or small companies who are not risking mission-critical operations on non-commercially supported privacy and security solutions because of the lack of accountability.

A healthcare provider might require guardrails that are deeply integrated with patient data systems and HIPAA compliance rules, while a financial institution may need guardrails that dynamically adapt to changing market regulations and customer privacy preferences. This lack of agility and customization hampers the effective implementation of trustworthy AI practices and poses potential risks to data privacy, security, and regulatory compliance. As AI systems become more deeply embedded in critical business processes, there is an urgent need for guardrail solutions that can be easily tailored to match each company's specific needs, especially in terms of privacy, security, and seamless integration with diverse data sources.

From a technical standpoint, it may seem that there are various approaches for mitigating possible issues in both model inputs and outputs. One smaller issue is that most guardrails focus on the model output, while the user itself is a part of a conversation creating context. Contextual under-

standing is still a significant challenge since different interpretations of what constitutes appropriate behavior are tied not only to language pragmatics but also to societal norms and cultural contexts that differ in specific cases, along with usually inaccessible individual preferences. Various static or too technical rules can lead to insufficient responses and further mistrust in AI systems.

A second issue is a much larger one: what kind of AI ethics are we talking about here? First, fine-tuning the model already enters a certain kind of *ethical bias*, by encouraging or discouraging certain behavior at will. Second, most users are not technical enough to allow for their values to be used as ethical guidance for the large language model they are using for their personal or business purposes. Various guardrails and ethical fine-tuning of models can inadvertently reinforce existing biases which can lead to perpetuating stereotypes or participating in further discrimination of certain groups. Lack of user awareness and understanding regarding the black-box-like nature of how guardrails operate can lead to more mistrust in AI systems.

The biggest problem here is again an ethical one, which is, paradoxically, not really emphasized in a problem resolving an issue in AI ethics. Namely, determining what constitutes ethical behavior is inherently complex, and can be tied to the idea of ethical pluralism that can allow for several values that may be equally correct and in contradiction, i.e., that there are many different moral values [37]. On one hand, such an approach recognizes the diversity of values across different individuals and cultures, promotes tolerance, and highlights the complexity of ethical dilemmas. On the other hand, it is easy to fall into a trap of complete relativism, by considering all positions equally

morally valid.

We will not try to resolve whether there is a lack of universality of moral principles or standards, but we will see that potential clashing and difficulty in the resolution of moral dilemmas, along with the possibility of incoherent ethical judgments may only be seen as a meta-problem. That is, overall, moral values may be in contradiction when analyzed from a superset perspective, but very rarely one will find practical examples of contradictions between the end user's ethical values when applicable to AI guardrails. If we truly want an exercise in practical ethics, the user needs to be able to configure its own guardrails that emphasize the following:

1. **Promotion of ethical autonomy.** Recognizing that ethical decisions often involve subjective considerations that vary not only between organizations that use AI systems but between end users as well.
2. **Enhanced transparency.** Providing the user with the ability to configure guardrails enhances transparency which is often an issue with the black-box system against which XAI is acting.
3. **Continuous improvement.** Users can provide feedback and refinement of AI systems' ethical frameworks that can be used not only by developers but organizations themselves to quickly improve the ethical robustness of their AI systems.
4. **Organizational alignment.** The concept of *aligned AI* is too general for many end users' needs since various organizations and individuals have distinct guidelines that govern their acts and operations.

5. **Contextual pragmatics.** Different contexts require different ethical considerations since what is appropriate in a healthcare setting may be completely different from what is appropriate in a finance sector.

4 Ethically-compliant guardrail design

4.1 Policies and rules

In modern conversational AI systems, ensuring compliance with rules and regulations is crucial to establish and maintain not only legal requirements but ethical standards that an organization or an individual wishes to maintain. The proliferation of AI assistants and the usage of conversational AI across different domains and disciplines necessitates robust yet user-friendly mechanisms to enforce ethical standards and rules governing content and behavior.

The proposed architecture organizes various types of *rules* into *policies*, allowing for structured but easily configurable enforcement within customizable *AI assistants*, that consist of a combination of policies. There are two ethical key points here. The first one is that a set of rules can be pre-built by the organization or the service provider, and the second one is that the user can add or modify further rules in order to fully customize an ethical guardrail.

The end user of such assistants can be individual users or organizations that care about AI safety. Since LLMs learn and use the data they have been fed as well, issues in privacy and security arise as a valid ethical concern in the usage of various large language models. The purpose of such rules is

to prevent any sensitive or unwanted data from reaching the LLM providers and to prevent any such data from reaching the end user as well.

4.2 Types of rules

There are three main types of rules that differ in their technical difficulty and in their strength. The user can use all or some of these to create fully customizable policies reflecting their ethical choices and privacy concerns.

The first one comprises *static rules* that consist of predefined patterns that AI assistants use to identify and filter easily predictable sensitive information such as email addresses, social security numbers, phone numbers, and other *personally identifiable information* (PII). For example, a regex or a similar natural language-processing (NLP) pattern recognition mechanism might detect or mask out PII to prevent unintentional exposure of private data to third-party LLM providers.

Natural-language rules are expressed by the user in human-readable language and provide guidelines on what behavior and content to either encourage or avoid in conversations made with an LLM. For example, “never mention any content inappropriate for children below the age of 12” or “avoid conversation about religion“. This type of rule is fully customizable and can cover a wide range of considerations, either from maintaining a polite discourse and avoiding offensive language to adhering to industry-specific regulations. These are different from system prompts¹⁴ since they can be combined with them without altering the original system prompt. System

¹⁴System prompts guide the way AI models interpret and respond to user queries. See e.g. [38] for more details.

prompts also refer to the AI output only, but natural-language rules can be used for user input as well.

In order not to fall in the same trap of using a LLM to mitigate responses from an LLM, two approaches can be taken here. First, a natural language rule can be enforced using natural language processing techniques without using LLMs. For example, various predefined lists of keywords and phrases can be used along with the user-defined description so that such lists can be avoided, along with various types of sentiment analyses and lexicon-based approaches. Another option is for an organization or a more technical user to host their own open-source versions of large language models such as Llama or Mixtral, leaving all of the user’s data in their own hands.

Trained classifier rules utilize machine-learning models to involve classifiers trained on labeled datasets. The user can start creating their own dataset by adding examples that are to be denied by the classifier and ones that are to be allowed. The user can upload their own datasets, use publicly available ones for fine-tuning, or use LLMs to generate synthetic few-shot examples that are similar to the ones chosen by the users. For instance, a classifier trained to detect urgency in medical assistance could prioritize responses to critical medical inquiries over general queries.

Non-technical users are likely to find natural-language rules particularly appealing due to their simplicity and accessibility. These rules allow users to express their ethical guidelines and preferences in plain, everyday language without needing to understand complex technical details or programming. For example, a business owner might easily set up rules like “ensure all communications remain professional” or “avoid discussing sensitive topics

like politics and religion”, without requiring any specialized knowledge of natural language processing or machine learning. However, specific trained classifiers, when provided with enough data, can offer a more nuanced and precise level of control over AI behavior. These classifiers can be tailored to recognize and respond to highly specialized or context-specific content, making them particularly valuable in industries where precision is critical.

4.3 Policies

Various rules can be combined into *policies*. Rules within policies are evaluated sequentially by default to determine compliance. Static rules come first so that immediate checks based on easily predictable patterns can be detected upfront and filter out sensitive information. Next, natural-language rules are then enforced, influencing the conversation that aligns with the user-defined ethical norms and requirements. Lastly, trained classifier rules classify both inputs and outputs based on learned categories, which allows a more nuanced understanding and response generation.

This default sequence can be altered by the user to create a *hierarchical chain of rules*, again emphasizing various ways a compliant system might be customized that depend on the end user’s privacy and ethical preferences. All of the policies can be used for both the AI output and the user’s input.

4.4 Assistants

A combination of input and output policies, along with preferred system prompts and action items creates an *AI assistant*. An *input policy* governs how the user can behave and what questions they can ask, ensuring compli-

ance with the organization’s values and regulations. An *output policy* follows the usual AI guardrail goal: govern how the AI assistant responds and interacts with its users, which ensures that responses align with user preferences or organizational standards.

When a rule within a policy is violated, this is labeled as inappropriate content so that the system can trigger a corresponding action, customized by the user. The user can choose what happens if a policy is violated, i.e., if a breach in one or more rules is detected. First, a redaction might happen and the interaction between the AI system and the user may continue seamlessly, but the sensitive data is never sent to a third-party LLM provider. For example, a rule may detect the user providing their social security number or trade secrets, and such information may be redacted. Second, a blocking violation may occur, in which the interaction is completely stopped, and, in case of organizations, there may be additional actions happening, such as sending a warning notification, logging the breach, or notifying a human in the loop. A warning message is sent to the user, informing them of the policy violation and providing context-specific feedback or instructions.

Depending on the severity of the violation, the AI assistant may restrict further interaction until the issue is resolved or avoided. For instance, in cases of repeated or severe policy breaches, the assistant could be programmed to temporarily block the user’s access, escalate the issue to a human moderator, or offer corrective guidance to ensure compliance moving forward. This dynamic and responsive approach to rule enforcement not only maintains the integrity of interactions but also fosters a safer and more controlled environment for all users involved.

Ultimately, the assistant’s ability to dynamically enforce these policies ensures that AI-driven interactions remain aligned with the ethical standards and operational goals of the end user or an organization, while also providing the flexibility needed to adapt to evolving requirements and contexts.

4.5 Framework summary

This architecture, as illustrated in Figure 1, provides a comprehensive framework for managing and enforcing rules in AI assistants, ensuring user trust and compliance across various domains. Deployment of AI guardrails (cf. [31], [36], [9]) represents a crucial step towards mitigating ethical concerns in AI systems.

However, these systems, never mind the good intentions behind them, often operate as black boxes to end-users, lacking transparency in how ethical decisions are made. In compliance with XAI, efforts have been made to enhance transparency in the guardrail portion of an AI system. The prototype in action is available publicly as an AI Trust Platform¹⁵ demonstrating these rules in practice.

5 Ethical pluralism in AI design

Ethical pluralism background acknowledges that diverse individuals and organizations may hold varying ethical perspectives and values, which underscores the need for customizable guardrails that can accommodate different ethical frameworks. By allowing the users to easily define their values incor-

¹⁵Preamble AI Trust Platform, app.preamble.com (2024).

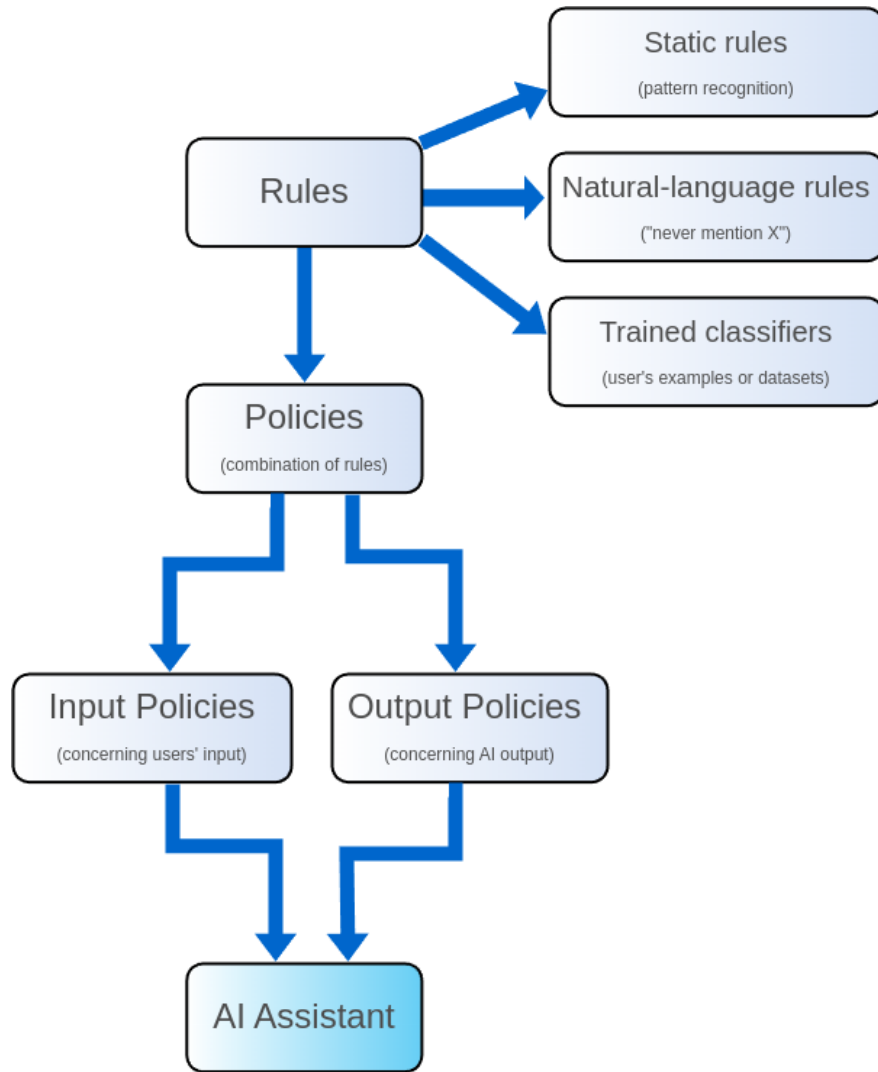


Figure 1: A hierarchical framework for managing and enforcing ethical standards in AI systems. This diagram illustrates the relationship between rules (static rules, trained classifiers, natural language rules), which combine to form policies. These policies are then integrated into AI assistants, ensuring that AI behavior aligns with specified ethical guidelines and operational goals.

porated into rules and policies based on their specific ethical concerns, an AI system can be fully aligned not only to an abstract moral model but with diverse standards and norms.

Russell’s [10] work argues that AI systems should be designed to align with human objectives (be “human compatible”), rather than aimlessly pursuing some predefined goals that may conflict with human values. Russell’s three principles include:

1. *The machine’s only objective is to maximize the realization of human preferences.*
2. *The machine is initially uncertain about what those preferences are.*
3. *The ultimate source of information about human preferences is human behavior.*

This approach is focused on the foundational aspects of AI ethics, while the guardrails approach in this paper emphasizes *practical* mechanism for implementing ethical standards in AI systems. The framework focuses on customizable rules and policies that allow the users to define ethical guidelines specific to their values and contexts, which acknowledges the diverse ethical perspectives that are too vague in Russell’s approach. The notion of “human preference” is already subject to questions of universality and relativism.

Both of these approaches emphasize *transparency*, calling out for clear explanations of AI behavior and reasoning, and by allowing users to configure and monitor ethical rules and policies, they can easily make ethical decisions or readily see the consequences of such rules in practice (with belief-revision

possibilities). A practical ethical AI agents should then follow a similar set of principles:

1. *AI system's objective is to maximize the realization of user-specified ethical guidelines and values.*
2. *The machine and user are initially uncertain about what those preferences are, but the user has to have a transparent yet fully configurable way to communicate such preferences.*
3. *The ultimate source of information about user's preferences is a set of rules and policies curated by the user, that can be analyzed and revised at will, according to the user's needs.*

The user-centric ethical design of AI guardrails is pivotal in navigating the complexities of addressing the value pluralism and the overall mistrust in AI systems that motivated the desire for explainable AI.

6 Possible conflicts

One possible issue is the question of conflicts between various guardrails inside a described framework. In addressing the complexities of guardrail conflicts within AI ethics systems, it's crucial to recognize the diverse scenarios in which such conflicts may arise and the mechanisms by which they can be managed. In this section, we will explore various cases of guardrail opposition, categorize them based on the nature of their conflict, and propose various strategies for their resolution.

6.1 Case 1: Complete and Permanent Opposition

In this scenario, Guardrail A and Guardrail B are always in complete opposition, meaning their ethical vectors as policy combination representations are exact opposites, with a dot product of -1 . This kind of conflict represents a situation where two ethical directives are fundamentally irreconcilable. For instance, one guardrail might prioritize absolute privacy, while another emphasizes full transparency. Such a scenario could be identified through static analysis, flagging the inherent opposition before deployment, by analyzing the mathematical relationships or logical conditions set by these policies.

Variant I: If Guardrails A and B are the only active guardrails, their mutual negation leaves the system without any ethical direction, resulting in an “ethically blind” state. This condition is particularly problematic as it disables the AI’s ethical guidance entirely.

Variant II: When other guardrails are active alongside A and B, the system may still function ethically as intended, relying on these additional guardrails. However, the persistent opposition between A and B may cause inconsistencies in ethical reasoning.

Variant III: If all active guardrails are engaged in mutual negation, the system is left without any moral guidance, despite having multiple guardrails. This total negation creates a scenario where the AI operates without ethical constraints, which is highly undesirable.

6.2 Case 2: Permanent but Limited Disagreement

Here, Guardrails A and B are generally opposed but not completely, with a dot product close to -1 , such as -0.9 . This situation mirrors the persistent yet manageable disagreements seen in political discourse, such as those between major political parties. Although there is opposition, it allows for some degree of consensus through weighted averaging. Static analysis could identify these cases, enabling adjustments to achieve a balanced ethical stance.

6.3 Case 3: Conditional Opposition

In this case, Guardrails A and B are only sometimes in complete opposition, i.e., their dot product is sometimes -1 , depending on the specific context or input. The ethical vectors might align in some situations but conflict in others.

Variants I, III, and III from Case 1 apply here as well.

6.4 Case 4: Conditional but Limited Disagreement

In this case, Guardrails A and B sometimes have opposing values but to a lesser degree, with a dot product around -0.9 . This situation is akin to a temporary political disagreement, where opposition exists but does not preclude finding common ground. Such conflicts are usually manageable within the system's existing framework, using weighted averaging to navigate the disagreement.

6.5 Conflict Resolution Strategies

To manage these conflicts, we consider the following strategies:

1. **Weighted Averaging System.** In most scenarios, especially in Cases 2 and 4, a weighted averaging system allows for nuanced ethical reasoning by considering the strengths of various guardrails. However, this approach struggles with Cases 1 and 3, where complete opposition can lead to ethical paralysis.
2. **Strict Order or Hierarchy of Precedence.** To address the limitations of weighted averaging, a strict order of precedence can be established, where higher-priority guardrails override others in cases of conflict. While this prevents total ethical blindness, it risks allowing weakly-held opinions from high-precedence guardrails to dominate more strongly-held positions from others, leading to potential ethical imbalances.
3. **Hybrid Approach: Conditional Precedence.** A hybrid approach could involve using weighted averaging as the default method but reverting to a strict order of precedence when mutual negation is detected (e.g., in Case 1/I, 1/III, 3/I, or 3/I). This system would handle temporary conflicts effectively while ensuring that permanent oppositions trigger alerts, warning users that the system is operating under constrained ethical guidance.
4. **Contextual Triggering.** The AI system can be designed to apply different guardrails based on the specific context or scenario. For

example, in situations involving sensitive personal information, the privacy guardrail might be activated, overriding the transparency directive. Conversely, in situations requiring public accountability, the transparency guardrail could take precedence. This approach allows for a dynamic resolution of conflicts based on real-time analysis of the situation.

5. **User resolution.** In cases where automatic resolution is challenging or where both guardrails are of equal importance, the system could flag the conflict for human intervention. Users or administrators can then manually decide which guardrail should take precedence in the given situation. This method is particularly useful in high-stakes environments where nuanced human judgment is necessary, invoking the *human-in-the-loop* strategy often present in AI systems.

Understanding and addressing guardrail conflicts is essential for the development of robust AI systems. While Cases 2 and 4 represent manageable, everyday ethical disagreements, Cases 1 and 3 pose significant challenges that require thoughtful design and conflict resolution strategies. By implementing a combination of weighted averaging and conditional precedence, we can create systems that maintain ethical integrity even in the face of complex, conflicting directives. This approach ensures that AI systems remain aligned with diverse human values, enhancing trust and reliability in their ethical behavior.

7 Final remarks

Our goal was to show that this approach of designing systems that prioritize user autonomy, transparency, and continuous improvement is a nudge in the right direction that promotes a collaborative approach where ethical standards evolve in response to feedback. Moving forward, continued research and innovation in AI ethics should aim to enhance transparency, accommodate diverse ethical perspectives, and allow the end-users to navigate the ethical aspect of AI technologies effectively.

The introduction of customizable guardrails not only enhances the ethical robustness of AI systems but also fosters greater transparency and trust. As AI systems become increasingly integrated into decision-making processes, the ability to understand and influence the ethical reasoning of these systems becomes paramount. The framework proposed in this paper aims to bridge the gap between abstract ethical principles and practical implementation, offering a structured yet adaptable approach that can evolve alongside technological advancements and societal changes.

Moreover, the strategies for resolving conflicts between guardrails shed more light on the complexity of ethical decision-making in AI systems. By providing a range of resolution mechanism, we ensure that AI systems can navigate ethical dilemmas in a manner that is both contextually appropriate and aligned with user expectations. This adaptability is particularly important in high-stakes environments where the consequences of AI decisions can have far-reaching impacts.

Looking forward, the continued refinement of these frameworks will be essential as AI systems are deployed in increasingly diverse and sensitive areas.

Future research should focus on expanding the capabilities of guardrail systems, integrating more sophisticated context-awareness, and exploring new ways to involve users in the ethical governance of AI. Additionally, ongoing collaboration between AI developers, ethicists, and end-users will be vital in ensuring that AI systems remain not only technically advanced but also ethically sound.

8 Declarations

The authors received no funding for this work.

References

- [1] V. C. Müller, “Ethics of Artificial Intelligence and Robotics,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta and U. Nodelman, eds.), Stanford University, Fall ed., 2023.
- [2] T. Baer, *Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists*. New York: Apress, 2019.
- [3] R. Lemos, “Employees Are Feeding Sensitive Biz Data to ChatGPT, Raising Security Fears.” <https://www.darkreading.com/cyber-risk/employees-feeding-sensitive-business-data-chatgpt-raising-security-fears>, 2023. DarkReading.com.
- [4] A. F. Winfield, K. Michael, J. Pitt, and V. Evers, “Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems

- [Scanning the Issue],” *Proceedings of the IEEE*, vol. 107, pp. 509–517, March 2019.
- [5] J. H. Moor, “The Nature, Importance, and Difficulty of Machine Ethics,” *IEEE Intelligent Systems*, vol. 21, pp. 18–21, Jul./Aug. 2006.
- [6] U. Agarwal, K. Tanmay, A. Khandelwal, and M. Choudhury, “Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language We Prompt Them in,” in *LREC-COLING 2024*, 2024.
- [7] H. J. Branch, J. R. Cefalu, J. McHugh, L. Hujer, A. Bahl, D. del Castillo Iglesias, R. Heichman, and R. Darwishi, “Evaluating the Susceptibility of Pre-Trained Language Models via Handcrafted Adversarial Examples,” 2022.
- [8] A. Etzioni and O. Etzioni, “Incorporating Ethics into Artificial Intelligence,” *The Journal of Ethics*, vol. 21, no. 4, pp. 403–418, 2017.
- [9] Meta.com, “Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations,” December 07 2023.
- [10] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking, 2019.
- [11] K. Siau and W. Wang, “Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI,” *Journal of Database Management (JDM)*, vol. 31, no. 2, p. 14, 2020.

- [12] J.-S. Gordon and S. Nyholm, “Ethics of Artificial Intelligence.” <https://iep.utm.edu/ethics-of-artificial-intelligence/>, 2024. Internet Encyclopedia of Philosophy.
- [13] J. R. Searle, “Minds, Brains, and Programs,” *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–424, 1980.
- [14] A. M. Turing, “Computing Machinery and Intelligence,” *Mind*, vol. 59, no. 236, pp. 433–460, 1950.
- [15] B. J. Copeland, “Artificial Intelligence.” <https://www.britannica.com/technology/artificial-intelligence>, 2020. Encyclopædia Britannica.
- [16] R. V. Yampolskiy, “AI-Complete, AI-Hard, or AI-Easy – Classification of Problems in AI,” *Artificial Intelligence Review*, vol. 42, no. 3, pp. 251–261, 2012.
- [17] K. Šekrst, “AI-Completeness: Using Deep Learning to Eliminate the Human Factor,” in *Guide to Deep Learning Basics* (S. Skansi, ed.), pp. 117–130, Cham: Springer, 2020.
- [18] S. Skansi, *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*. Springer, 2018.
- [19] OpenAI, “Better language models and their implications.” <https://openai.com/index/better-language-models/>, February 14 2019. OpenAI.com.

- [20] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, and S. Stumpf, “Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions,” *Information Fusion*, vol. 106, p. 102301, 2024.
- [21] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 4th ed., 2021.
- [22] R. Ngo, L. Chan, and S. Mindermann, “The Alignment Problem from a Deep Learning Perspective,” in *International Conference on Learning Representations*, 2022. arXiv:2209.00626.
- [23] R. Merritt, “What Is Retrieval-Augmented Generation, aka RAG?.” <https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/>, November 15 2023. NVIDIA.com.
- [24] Y. Dong, R. Mu, G. Jin, Y. Qi, J. Hu, X. Zhao, J. Meng, W. Ruan, and X. Huang, “Building Guardrails for Large Language Models,” *arXiv preprint arXiv:2306.07500*, 2023.
- [25] Y. Wang and L. Singh, “Adding Guardrails to Advanced Chatbots,” *arXiv preprint arXiv:2306.07500*, 2023.
- [26] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, “Instruction Tuning for Large Language Models: A Survey,” *arXiv preprint arXiv:2308.10792*, 2023.

- [27] L. P. Kaelbling, M. L. Littman, and A. W. Moore, “Reinforcement Learning: A Survey,” *arXiv preprint cs/9605103*, 1996.
- [28] Y. Lei, D. Ye, S. Shen, Y. Sui, T. Zhu, and W. Zhou, “New Challenges in Reinforcement Learning: A Survey of Security and Privacy,” *arXiv preprint arXiv:2310.10501*, 2023.
- [29] Y. Huang and Q. Zhu, “Deceptive Reinforcement Learning Under Adversarial Manipulations on Cost Signals,” in *International Conference on Decision and Game Theory for Security*, pp. 217–237, Springer, 2019.
- [30] V. Behzadan and A. Munir, “Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks,” in *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pp. 262–275, Springer, 2017.
- [31] T. Traian Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen, “NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails,” *arXiv preprint arXiv:2310.10501*, 2023.
- [32] HuggingFace, “all-minilm-l6-v2.” <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>, 2024.
- [33] T. Cover and P. Hart, “Nearest Neighbor Pattern Classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [34] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, “Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment,” *arXiv preprint arXiv:2312.12148*, 2023.

- [35] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, and T. D. et al., “Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations,” *arXiv preprint arXiv:2312.06674*, 2023.
- [36] GuardrailsAI.com, “Guardrails AI.” <https://www.guardrailsai.com/>, Accessed: 2024.
- [37] E. Mason, “Value Pluralism,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta and U. Nodelman, eds.), Metaphysics Research Lab, Stanford University, summer 2023 ed., 2023.
- [38] M. Zheng, J. Pei, and D. Jurgens, “Is ”A Helpful Assistant” the Best Role for Large Language Models? A Systematic Evaluation of Social Roles in System Prompts,” *arXiv preprint arXiv:2311.10054*, 2023.