The Assessing Risks and Impacts of AI (ARIA) Program Evaluation Design Document

Reva Schwartz[1], Gabriella Waters[2,3], Razvan Amironesei[1], Craig Greenberg[1], Jon Fiscus[1], Patrick Hall[2,4], Anya Jones[2,3], Shomik Jain[2,5], Afzal Godil[1], Kristen Greene[1], Ted Jensen[1], and Noah Schulman[1]

[1]National Institute of Standards and Technology (NIST)
[2]NIST Associate
[3]Morgan State University
[4]Hall Research
[5]Massachusetts Institute of Technology

December 20, 2024

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Disclaimer

Certain commercial equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Contact Information
aria_inquiries@nist.gov

1. Introduction

As artificial intelligence (AI) technology becomes increasingly integrated into our daily lives, there is no shortage of difficult questions about the risks it poses and whether it will negatively or positively affect people and society. It is difficult to determine whether, how, and to whom AI models with advanced capabilities pose risks. AI actors require detailed information to understand risks and to decide whether to procure, deploy, or use AI in their specific contexts [1].

ARIA (Assessing Risks and Impacts of AI) is a NIST evaluation-driven research program to develop measurement methods that can account for AI's risks and impacts in the real world. The program establishes an experimentation environment to gather evidence about what happens when people use AI under controlled real-world conditions. In contrast to current approaches that rely on probabilities and predictions, ARIA will enable direct observation of AI system behaviors and potential impacts on users. ARIA pairs people with AI applications in scenario-based interactions designed around specific AI risks and studies the results. Applications are submitted to NIST from around the globe and are evaluated on the basis of whether risks materialized in the scenarios, and the magnitude and degree of resulting impacts. Participating teams will learn whether their applications can maintain functionality across the varying contexts of the test environment.

Figure 1 presents the process flow diagram of ARIA's experimentation environment that can enable a comprehensive view of AI risk and impacts before, during, and after they materialize in user interactions[2]. The environment's configurable design can provide an almost limitless number of simulations to fill in missing information about what actually happens when people use AI technology in the real world. Initially, ARIA will principally focus on risks that can be directly observed through user interactions with AI technology[3]. Over time, based on available resources and with input from the ARIA research community, the program may potentially expand to examine broader AI risks and related impacts, such as to the workforce.

Evaluation of AI applications starts in ARIA's three-level testbed, in which each level uses a different testing approach to explore potential risks and impacts:

1. Model testing: confirm claimed capabilities
2. Red teaming: stress test and attempt to induce risks
3. Field testing: examine positive and negative impacts that may arise under regular use

[1] AI actors are "those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI" [OECD (2019) Artificial Intelligence in Society—OECD iLibrary] For a full list of AI actors, see the NIST AI RMF.

[2] All tests with users will follow standard human subject protocols and receive approval from the NIST Research Protections Office (RPO) prior to enrolling human participants.

[3] NIST intends to initiate testing with the taxonomy of risks defined in the AI Risk Management Framework Generative AI Profile [1].

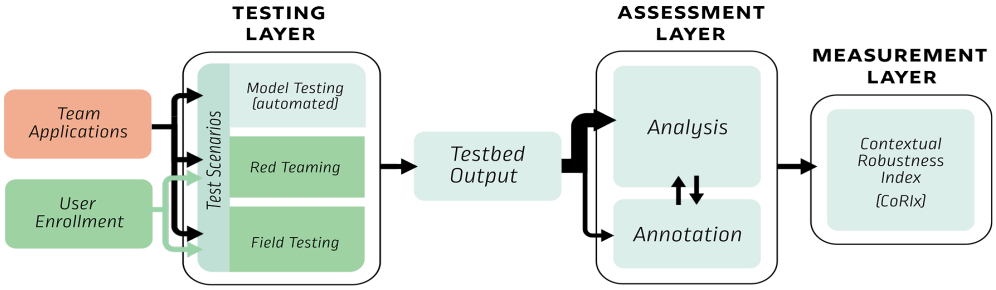

Fig. 1. ARIA's experimentation environment simulates deployment contexts to understand how, when, and for whom AI risks and impacts materialize.

Next, the testbed output for each application is annotated and analyzed in ARIA's assessment layer to determine whether risks materialized in the interaction and to characterize the outcomes. Finally, the application results are calculated in the measurement layer using the Contextual Robustness Index (or CoRIx), ARIA's measurement instrument and suite of metrics. The CoRIx measures whether AI applications can maintain robust and trustworthy functionality across deployment contexts.

Dialogues collected in the ARIA environment will be curated and anonymized and are planned to be publicly released after each evaluation series. The publication of ARIA's methods, metrics, practices and tools will facilitate adoption and scaling across industry and research settings. ARIA's CoRIx instrument and suite of metrics will also be collaboratively developed and released for broad adoption.

Selected ARIA metrology terminology are as follows:

- **Assessment:** Action of applying specific documented criteria to a specific software module, package or product for the purpose of determining acceptance or release of the software module, package or product [2].
- **Benchmarking:** (i) Standard against which results can be measured or assessed; (ii) Procedure, problem, or test that can be used to compare systems or components to each other or to a standard. [2]
- **Evaluation:** (i) Systematic determination of the extent to which an entity meets its specified criteria; (ii) Action that assesses the value of something [2].
- **Measurement:** (1) Quantitative measurement is the act or process of assigning a number or category to an entity to describe an attribute of that entity [2]. (2) Qualita-

tive measurement is based on descriptive data derived from observations, interviews, focus groups, or open-ended text fields in surveys.

2. Background

New and rigorous methods, metrics, processes, data, and skills are required to evolve the field of AI risk measurement and to account for what happens when people use AI in real world contexts.

2.1. Overview of NIST Evaluation-Driven Research

NIST's Information Technology Laboratory (ITL) has a long history of evaluation-driven research of technology. [4] For decades, ITL has hosted technology evaluations using a common set of tasks, data, metrics and measurement methods. This approach reduces overhead, enables reproducibility, and drives identification of the most promising research directions for technology improvements. NIST evaluations are long-term exploratory research efforts that span multiple iterations, tasks and challenge problems, and are open to all who have interest.

NIST's evaluation program outcomes can inform organizational decision making, but NIST does not define measurement thresholds, select or weigh in on which metrics or methods should be used by external evaluators or entities, or otherwise make recommendations about which technique is the best available for a given purpose. Evaluation plans specify requirements for expected application behavior and the methods to evaluate performance. Evaluation workshops are held at a regular cadence to collaboratively identify areas of research refinement and expansion, and to plan future evaluations.

For the ARIA program, NIST has developed and will supply the following:

- Infrastructure for testing, annotating and scoring submitted AI applications
- User recruitment, enrollment and management for red teaming and field testing activities
- Scripts for model testing tasks
- Submission criteria for AI applications
- Testing protocols, risk proxy scenarios, and risk guardrail criteria
- Annotation schema and process, annotators, and annotated outputs
- Scoring methodology, evaluation outcomes and reporting

2.2. Strengthening AI Risk Measurement

Performance and risk in AI systems are distinct yet interconnected aspects that require separate consideration in measurement paradigms. Measures of accurate performance may

[4] For more information about NIST AI technology evaluations see: https://www.nist.gov/programs-projects/ai-measurement-and-evaluation/nist-ai-measurement-and-evaluation-projects.

indicate that a system can accomplish tasks effectively, but it does not necessarily correlate with lower risk likelihood or positive outcomes in the real world. An AI system can excel in its intended functions while still contributing to undesired or negative impacts. ARIA enables a broader view of the orthogonality between performance and risk by examining AI in context. The three-level testbed is designed to surface AI's technical capabilities, potential for risk, and resulting impacts under different test scenarios. Dialogue output from the three levels are annotated in the same way to reveal salient differences and to provide a more complete picture of how AI risk arises in the real world.

NIST's recent efforts in the fields of AI risk management, generative AI risk, and trustworthy and responsible AI serve a foundational role for ARIA. For example, organizations implementing the AI Risk Management Framework's (AI RMF) [3] Measure function can expect to benefit from the Test, Evaluation, Verification and Validation (TEVV) methods and related AI risk measurement practices developed in ARIA. In particular, ARIA's risk focus is directly informed by the framework, which defines risk as:

> "***The composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats***." (AI RMF page 4)

Current AI evaluation approaches are limited in several ways:

- AI risk measurement requires detailed information about which risks and impacts actually materialize in the real world, how, and for whom. Without this information, validating the accuracy of AI risk estimates is difficult.
- Performance-based metrics, such as accuracy, are insufficient for assessing the type, magnitude or degree of AI's impacts in the real world.

ARIA will collect data that can corroborate what happens when people use AI in the real world.

ARIA aims to simulate the conditions of the real world to build up methods that can:

- provide missing detail about how people use and interact with AI technology in deployment settings;
- detect whether and how often risks materialize in real world settings, and who may be impacted;
- characterize and categorize AI risks and impacts to people;
- link specific AI risks to specific impacts;
- gauge whether risk and impact mitigation approaches achieve their intended goals; and

- assess AI system trustworthiness, explore the tradeoffs between trustworthy characteristics, and inform the design and development of trustworthy AI.

Table 1. Summary of comparisons between current approaches and ARIA testbed-style experiments

Aspect	**Traditional AI Evaluations**	**ARIA Experiments**
Focus of assessment	Model capabilities and performance	Materialized risks, magnitude and degree of impacts, AI trustworthiness
Evaluation approach	Model-centric, benchmarking	Human-centric, user interactions with AI applications in context-specific scenarios
Measurement approach	Primarily quantitative	Mixed methods (combination of quantitative and qualitative techniques)
Metrics	Primarily accuracy	Contextual robustness (suite of metrics)
Stakeholder community	AI actors on the lifecycle - primarily development teams	Operators, end users and potentially impacted individuals; Deployers; Subject matter experts; Metrologists
Test environment	Built for each vertical domain	Horizontal/Multipurpose

3. Motivating Factors for the ARIA Program

This section describes some key challenges in AI risk measurement that motivated ARIA's design and research focus to date. Subsection 3.1 describes current limitations when measuring AI system risks and impacts that manifest in real-world settings. Subsection 3.2 describes measurement challenges due to the broad variability in how people may use and repurpose AI systems.

3.1. Risk Measurement Requires Different Methods and Data Than Performance-Based Measurement.

Risks can arise from how AI systems process, generate, and disseminate information in real-world settings. These risks go beyond model accuracy and may include harmful bias, difficulty controlling public exposure to dangerous, violent, and harmful content, and leakage or unauthorized disclosure of private data. [5] Many of these risks and resulting impacts can materialize via interactions among the AI system, users, and the broader socio-technical environment [4, 5].

Like risk estimates for any field, AI risk estimates use statistical models that are limited by the factors included or represented in that model. To effectively capture detailed insights about risk, measurement methods need to contextualize factors beyond the technical AI system itself and consider the following:

- **Corroborating data:** Did a risk and resulting impact materialize in context?
- **Detail about the consequences:** Can the risk and resulting impact be contextualized within the relevant setting?
- **Effectiveness of existing controls:** Do risk mitigation approaches achieve their intended real-world aims?
- **Generalizability:** Can the statistical model be applied to similar risks in other real-world settings?

For the purposes of the ARIA program the following definitions are provided to foster a shared understanding of **context**.

[5] For more information about risks related to generative AI, see Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile [1].

Context: The parameters in which interrelated factors, purposes, and circumstances may shape individual and collective perceptions, interpretations, and expectations about the functionality and impacts of AI technology, and resulting actions.	**Context-of-use:** "Comprises a combination of users, goals, tasks, resources, and the technical, physical and social, cultural and organizational environments in which a system, product or service is used[; ...] can include the interactions and interdependencies between the object of interest and other systems, products or services." [6]	**Contextualization:** Act of placing a materialized risk within a broader setting for interpretation of its impacts and estimate its likelihood.

Real-World Conditions Can Be Integrated into Risk Measurement

Knowing whether an AI system produces a given output is only the first step in a chain of events about AI risks and impacts. At a minimum, it is also necessary to know whether individuals will actually be exposed to a potentially positive or negative impact; whether or how they will perceive, make sense of, and use or act on that output; and what the resulting effect can be. Capturing this type of information is currently difficult. For example, although AI users are often able to report incidents and contest AI-based decisions, there are few methods to conceptually and practically place those incidents in context.

Currently, AI risk probabilities are based on technical and physical factors of the AI model. AI practitioners evaluate model capabilities through computer simulations with benchmark datasets representing some aspect of the real world [7, 8]. Benchmark datasets are typically large-scale, static and retrospective, and focus on model performance in isolation. This type of evaluation approach is invaluable during model development but is unable to account for the benefits and risks of an entire AI application within its operational setting, that is, where risks typically play out in complex interactions with people. For example, some operational factors, such as human-AI configuration, secondary data use, or security incidents, can contribute to negative or positive outcomes, but are not typically included in today's risk estimates.

Baseline risk assessment methods, which often rely on theoretical models or limited testing suites, are also unable to systematically account for impacts in deployment settings. For small-scale (often singular) AI risk and impact assessments, it is difficult to map failure modes to specific AI system behavior without contextual information, or to tease out whether an outcome will be positive or negative once an AI system is deployed [9][10]. This type of testing can be particularly laborious given the broad and growing complexity of AI risks and resulting impacts.

Other challenges include data contamination, which occurs when AI models are exposed

to the data used in testing before the actual test [11], and task contamination, which occurs when models are exposed to examples that are similar to the benchmark task prior to testing [12]. Data and task contamination can lead to erroneously high performance estimates for computational benchmark results and the potential compromise of publicly available benchmarks when many AI systems share the same contaminated training data.

Use Case: Challenges Posed by Underspecification
A key limitation of benchmarks is underspecification, a phenomenon in machine learning when, for a model to perform well in the real world, technical best practices are not enough, and therefore additional domain expertise is required [13]. Sentiment analysis, a process used to categorize text by affective valence, provides a useful example of "underspecification" when the analysis output fails to encode sufficient domain expertise to reliably handle nuances in complex text data. [6] Language models have improved performance of sentiment classifiers, but challenges still persist for non-English or mixed-language text, sarcasm, and negation (which reverses the meaning of the text), and for other common speech and text-based behavior included in machine learning (ML) training data [14–16]. More generally, non-surface factors such as the meaning of language may be too nuanced and complex for sentiment modeling approaches to capture or place into context. Human-performed annotation and labeling can account for the contextual meaning of human communication, improve performance and address underspecification. ARIA's human-performed annotation processes aim to capture a wider variety of signals in user interactions with AI applications. These methods will start out as entirely manual and adapt over time to be less resource-intensive.

AI Lifecycle Practices Can Inform Risk Measurement
AI risks can materialize across the entire AI lifecycle, [7] as well as adapt and evolve over time and across conditions.

AI systems are not only technical artifacts constructed from data, compute, and algorithms, but also a product of the people who design, develop, deploy, and use them [18, 19]. These human factors are typically unavailable or abstracted away in AI lifecycle processes. AI models themselves are discrete representations of the real world and unable to account for real-time dynamics [13]. The various human behavioral and contextual factors in the datasets that underlie model development and engineering are further "flattened" by ML processes. This mismatch between the real world and the inner workings of AI applications often means that they can typically excel in design and development settings, only to struggle when faced with the noisy conditions of real-world deployment [13, 20].

Contextual information is difficult to integrate across the AI lifecycle. ML uses "lightweight" process-based techniques to facilitate technology production and scalability [21]. Contex-

[6] Real Toxicity Prompts is an example of a benchmark that relies on an ML toxicity classifier for scoring [17].
[7] ARIA uses the AI lifecycle described in the AI RMF [3].

tual information, however, can be thought of as a form of "thick description."[8] The elicitation, capture, processing, analysis, and integration of contextual information requires specific domain expertise and is customarily "slow" and difficult to scale. Various structured feedback mechanisms can be used to support contextually informed processes. However, a long-standing challenge in technology development is how to separate human-driven methods such as annotation or in-depth risk analysis from the individuals who conduct them, for scaling and automation. It is often difficult to determine which human tasks and processes are related to domain expertise, skills, and experience, and which can be packaged for automation. For example, processes that require domain knowledge may be either:

- automated and lose contextual detail along with awareness of what was "lost." or
- remain entirely manual and unable to deliver contextual benefits at scale.

Decisions across the lifecycle, like in any profession, can also be prone to subjective interpretation. AI actors may reach significantly different conclusions about salient risk factors–even for the same AI system [21, 23, 24], leading to erroneous decisions. For example, AI actors may opt to not deploy an AI system based on a risk that is not genuine.

Once released, AI models can be expected to skew in ways that are difficult for practitioners to predict and anticipate. For example, AI deployers consider expectations and needs of different user groups, [3] but they usually have limited visibility into other parts of the lifecycle and few opportunities to systematically get direct insights from people who use the technology. Evaluating in real-world testing conditions can improve understanding of the dynamic and multi-dimensional aspects of real-world risks, yet such approaches are currently nascent.

Currently, AI evaluators can provide contextual details across the AI lifecycle by:

- using dynamic documentation and transparency approaches;
- incorporating input from individuals with varied skills, perspectives, backgrounds and disciplines; and
- conducting participatory engagement with users and other stakeholders that are external to the organization.

Eliciting and capturing input from the public is itself a specific expertise that requires formal collection and analysis skills, and the ability to translate results to technical AI actors. [3].

Data That Can Corroborate Risk Are Required

One common approach to managing AI risk is to locate additional data related to the specific AI modeling tasks. However, additional data may not be useful if they lack detail

[8]Thick description is the ethnographic process of understanding actions in real-world context via the collection of in-depth qualitative data. For an understanding of thick description as it relates to technological systems, see [22].

about the materialized risk. In a quest for "corroborative" data, some AI researchers have taken a cue from cybersecurity efforts such as incident reporting, and red teaming. When applied to AI, incident reporting involves monitoring and tracking adverse AI events in deployment [25, 26]. These kinds of data can provide a sample of real-world AI risks but typically lack the necessary detail about how and for whom risks materialized. Adapted from cyber red teaming, AI red teaming practitioners test systems in realistic risk scenarios to produce adverse outcomes and to identify risk boundaries [27–29]. Depending on how they are designed, AI red teaming exercises can provide more experimental control over key variables to deduce how risks materialize[30]. However, because red teaming is adversarial in nature, it cannot–on its own–provide a complete picture of the AI risk surface or the positive impacts of AI technology.

Many organizations that deploy AI technologies capture and document real-world outcomes of system performance, along with user feedback directly from the deployment context. However, these data are not typically released for research purposes or collected with standard research or human subject protocols, limiting their availability and accessibility [31].

ARIA's experimentation environment is designed to capture materialized risk data and characterize the results. By simulating real-world conditions, the experimentation environment can also enhance the "testability" of AI risk measurement approaches and collect data to "corroborate" hypotheses about AI functionality in deployment contexts. As "the more empirical tests a theory passes the more valid it becomes" [32] ARIA can accelerate the development of valid and reliable tools, methods and data to drive the development of AI risk measurement science. The experimentation environment also provides a venue to falsify, refute, and revise ARIA's evaluation processes, enhancing their experimental validity, reproducibility, and generalizability [19, 33].

3.2. The Complexity of AI Use Cannot Be Overstated

Risk measurement is also complicated by the ambiguity, variability and heterogeneity of the social, cultural, and organizational contexts in which people use AI technology [34]. Users bring their own perspectives, expectations and mental models to their interactions with AI and interpretations of resulting outputs, all of which can vary widely. Individuals can reuse, misuse or re-purpose even the same AI system. For example, a health care worker using a hiring application may have different expectations about model outputs depending on the purpose, setting, and task, and interpret and act upon the resulting AI output in different ways [34]. User behavior and societal and cultural norms continuously adapt. An AI personal assistant might initially be perceived as helpful and trustworthy, but over time can reveal subtle negative impacts on users' decision-making skills or privacy.

The proliferation of Generative AI (GAI) technology illustrates the challenges posed by heterogeneity [34]. Prompt sensitivity and other dynamic factors in deployment require GAI models to adapt to user contexts. GAI systems, such as large language models

(LLMs), that are unable adapt to the communication nuances of user interactions and expectations in deployment may:

- Misinterpret user intent: Models may take sarcastic comments in a user prompt literally and produce inappropriate responses, or vice versa.
- Fail to adapt to dialogue tone: Models may produce overly casual language in a formal business setting, or vice versa.
- Ignore cultural nuance: Models may produce culturally insensitive remarks or recommendations to users.
- Mishandle sensitive topics: Models may frame controversial topics inappropriately.
- Provide irrelevant information: Models may provide generic responses that do not meet user specifications and requirements.
- Act overly familiar: Models may engage with users in a manner that clashes with user expectations for the setting.

Determining "correctness" or "appropriateness" of AI-generated output in a given setting could entail a potentially infinite set of "answers" from the prompter's perspective. Measuring this infinite set of responses is improbable. Notably, ARIA's experimentation environment does not seek to model a priori every possible risk associated with every option for how people interact with AI applications. Rather, the environment makes it possible to monitor whether a risk materialized in people's interactions with AI, and to explore the contextual aspects of that risk and resulting impacts when it does. In effect, the ARIA environment may sharpen collective knowledge about the conditions under which AI's opportunities and threats may be more likely to occur. This knowledge can improve modeling and algorithmic techniques for risk identification, classification, and assessment.

Use Case: Challenges Posed by Complexity–Keyword Filtering
Keyword filtering is commonly used to assess whether different types of content are present in the datasets underlying the models used in AI. Yet, keywords remain a blunt tool for capturing the complex nuances of people's language use in real-world interactions with AI. For example, AI actors may filter by keyword to block use of the word "death" when moderating discussion of sensitive topics [35]. This approach can fail to distinguish between benign uses of the word (e.g., "the death of a star") and genuinely sensitive uses (e.g., discussion of suicide). Keywords may also incorrectly flag the use of terms that differ based on cultural and societal norms - a challenge that is particularly difficult to address with the multipurpose nature of LLMs and their growing global use. Keyword-based approaches can also gloss over contextual factors such as user intent and situational appropriateness. These limitations may lead to people being overexposed to undesired information and not presented with the information they are seeking.

A set of practices for producing safer AI, reinforcement learning with human feedback

(RLHF) seeks to tune AI output to be less harmful by using human judgments [36, 37]. The process frames generated output within a narrow range of prespecified constraints and norms for the individual judgments, such as "helpful" and"harmful". Reducing the complexity and heterogeneity of the deployment context–and accepted model performance within it–to a narrow set of judgments creates a tradeoff. Although it is more predictable and less resource intensive than highly contextual annotation processes, the tuning process itself can obfuscate risk and lead to unintended consequences [38, 39].

Tuning and alignment approaches have similar limitations to other performance-based approaches:

- Generalizability: Individual opinions can differ greatly on the meaning of "helpful" and "harmful" based on their circumstances, preferences, expectations and aims when using the technology, along with culture, language, background and personal values. Because an AI-generated response deemed helpful in one context could be inappropriate or even dangerous in another, it is difficult to scale or generalize from these approaches [40].
- Reductionist: A narrow classification of human expectations and behavior cannot effectively account for the dynamics of real world deployment.
- Temporality: An AI action that appears helpful in the short term may have negative long-term effects that are not immediately apparent in tuning processes.

4. ARIA's Experimentation Environment

To build up AI risk measurement science, ARIA will gather evidence about whether risks actually materialize in simulated real-world conditions and characterize the resulting impacts. This section provides detailed descriptions of ARIA's testing, assessment, and measurement layers and how ARIA's evaluations will be conducted. The section uses an illustrative example of LLMs, but ARIA evaluations will be configurable and can be designed for different types of AI (predictive, generative), users (adversarial, everyday, domain experts), and risks.

4.1. Testing Layer

ARIA's three-level testbed is designed to elicit, capture and account for contextual nuance and variation in user interactions with AI applications. Testing layer requirements, scenarios and criteria for each ARIA evaluation series are described in evaluation plans [41]. The ability of ARIA testing layer processes to effectively simulate real-world conditions will be assessed after each evaluation series. This section describes the ARIA testbed, processes, and related material.

4.1.1. Testbed

The ARIA testbed is designed to generate large numbers of user AI interactions and output data.

1. Model testing: Fully automated, model testing is designed to confirm the claimed capabilities and limitations of the submitted application. The output of this level is the dialogues resulting from the responses to the automated prompts.
2. Red teaming: Red teaming entails different types of users who adversarially interact with AI applications to induce violation of guardrails that may potentially manifest risks and attempt to induce risks. The output of this level includes dialogues between red teamers and AI applications, post-session questionnaires, and listed red team strategies and outcomes for each scenario.
3. Field testing: Field testing entails individuals customarily interacting with AI applications to complete a specific task under conventional settings [9] The output of this level includes dialogues between field testers and AI applications, and post-session questionnaires.

Importing technology components for testing has consistently been a challenge for machine learning evaluations. As AI systems rely increasingly on large foundation models, importing models and a full algorithm to checkpoint an application's state for controlled testing becomes untenable. Further, ARIA's need to access a large test subject pool and

[9] This may include everyday use of technology such as AI-powered chatbots, writing assistants, or navigation systems for their designed purpose in either a professional or personal capacity.

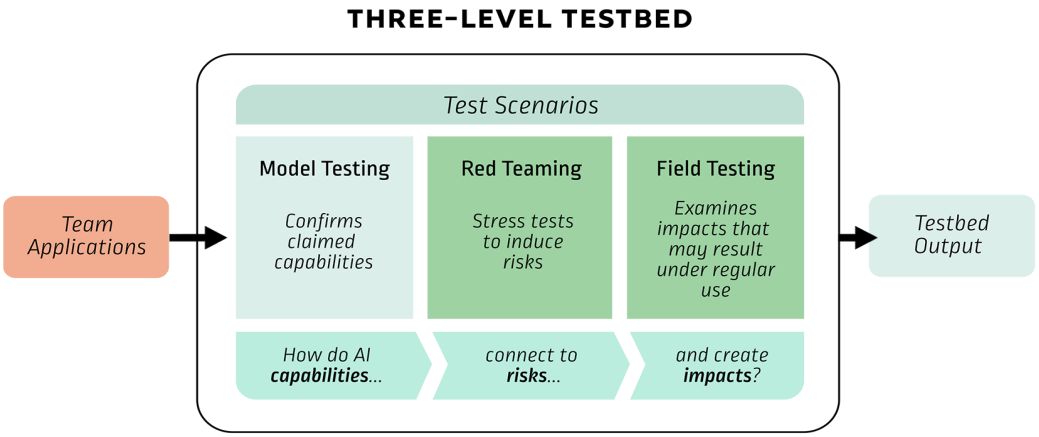

Fig. 2. Each level of ARIA's three-level testbed focuses on a different aspect of AI risk measurement.

to test systems over the Internet via a consistent User Interface, requires a re-imagining of the delivery process. ARIA's approach is to use a light-weight, Python Abstract Class that specifies a consistent set of dialogue interactions between an Internet accessible model and the user. To use this delivery method, participating model teams are required to provide NIST-credentialed Internet access to their model, whether public foundation model or self-hosted model. Participants also have the option to implement scenario guardrails within the Python Class or at their site.

The NIST testing harness will use a single user interface that hides the identity of the application/model under test, enabling an apples-to-apples comparison environment to the extent possible for the red teamers and field testers.

4.1.1.1. Testing stimuli and processes

Large-scale experiments require a set of shared tasks or scenarios to enable testing on the same "test track." ARIA facilitates this process through different test stimuli and related materials, described below.

Scenarios The ARIA framework employs pre-designed scenarios to assess key risks associated with AI usage. [10] While automated scripts execute scenarios at the model testing level, the red teaming and field testing process involve users interacting with AI applications through prompts guided by scenario-specific instructions. ARIA's scenarios are crafted to

[10]For more information about proxy scenarios used in the ARIA 0.1 pilot, see the Evaluation Plan [41].

isolate specific risk structures,[11], facilitate easy deployment, and support thorough, iterative exploration. Developing these test scenarios is a complex, time-consuming process that demands interdisciplinary expertise. As the ARIA community grows, it will collaboratively build extensive libraries of generalizable risk scenarios that benefit the broader AI evaluation community and enhance our understanding of AI-related risks in various contexts.

Evaluating real-world risks and impacts requires the ability to know when a risk actually materialized in the scenarios. ARIA uses a process[12] to approximate AI model guardrails and define permitted and prohibited application behavior in the testbed interactions for each risk scenario. Any violation of the application guardrails constitutes a "materialized risk" for deeper exploration. A guardrail is "violated" when the application exhibits one or both of the following prohibited behaviors:

1. Prohibited content is released
2. Permitted content is withheld

Instructions to users User instructions aim to ensure that red teamers and field testers follow the test scenarios as closely as possible during their interactions with the AI application. The instructions are designed for experimental control and to maintain external validity. Testers are provided the instructions prior to their interactions, to set forth a consistent understanding of their task for each scenario, and allowing them to approach their respective interactions as naturally as possible. Instructions are designed to be intentionally brief while allowing room to approach the scenario interactions in different ways.

Model testing scripts Model testing is designed to confirm the claimed capabilities and limitations of the submitted application. Unlike common benchmarking methods and efforts [7], [8], [42], [43], [44] that rely primarily on pre-prepared, static data sets designed to measure model accuracy, the scope of ARIA model testing is limited to risks. Two types of confirmations are used for testing the applications. First, the testing demonstrate that the application can generate content relevant to the scenario and can perform a useful function rather than generating safe but unusable content. Second, the testing also demonstrates that the application guardrails prevent production of violative content. Notably, ARIA model testing is less exhaustive than that in traditional model evaluations because its focus is on risk rather than performance-based metrics such as accuracy.

Questionnaires Questionnaires are used in the red teaming and field testing levels of the testbed to gain insights into user perceptions of risk and impact exposure and related information. Questionnaire items are related to perceptions of application output, task-specific impacts, and future behavior. Red teamer questionnaires also include a space for input

[11]The first ARIA test, a pilot study, investigates risks associated with generative AI. The risks are selected from NIST AI 600-1 "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile" [1].

[12]Called "test packets".

NIST Proxy Scenarios

A proxy scenario is one that is analogous, but not equal to, the actual scenario in which an AI system will eventually be used. NIST has successfully used proxy scenarios in scientific measurement efforts for decades. Proxy scenarios enable:

- repeatability and measurement consistency as compared to post-deployment observational approaches or benchmarking of AI system attributes, and
- apples-to-apples comparisons of different systems via a common set of challenge problems, tasks and measurement indices that are applicable to all.

Proxies in NIST evaluation protocols

Several NIST evaluations follow a similar, three-step, high-level process to utilize proxy scenarios:

1. *Define the use case(s) and proxy scenario(s).* To facilitate the development and reuse of evaluation tasks, NIST designs scenarios that mimic the real challenges, while ensuring accordance with practical and ethical constraints.
2. *Extract key enabling capabilities.* Instead of setting up tasks under every possible condition, NIST researchers identify relevant foundational variables to enable deeper investigation and create a reusable scenario that can be implemented under controlled conditions.
3. *Design experiments.* NIST provides methods, datasets, and measurements designed to shed light on the relevant foundational variables and address the real challenges under focus.

NIST has followed this similar process for various biometrics, information retrieval, and video and image analytics. To foster understanding of the use of proxies in evaluation protocols, two example proxy scenarios are provided below.

Proxy Scenario Example–NIST SRE: As part of the Speaker Recognition Evaluation (SRE) series, which has occurred regularly since 1996, NIST provides a common framework to enable the research community's scientific exploration of promising new ideas in the field. (i) The SREs have driven advancements in identifying speakers within conversational telephone speech recordings for real-world tasks such as user access to banking via voice, call center fraud detection, and personalization of voice-activated personal assistants (e.g, Siri and Alexa).(ii) Instead of setting up tasks directly in bank or call center conditions, NIST has focused on the general task of text-independent speaker recognition, introducing complex and broadly applicable real-world challenges into the evaluation. (iii) NIST has conducted experiments addressing a wide variety of factors in the SRE series, including signal and environmental noises, speech duration, vocal effort, language, and others.

Proxy Scenario Example–NIST ARIA: (i) Generative AI capabilities may pose risks to safeguarding privileged information such as private data, proprietary content and dangerous or classified information. A proxy scenario was designed around another type of privileged information, the TV plot spoiler. (ii) While the specific information of a TV spoiler is not a serious risk to the public, it enables a repeatable task that can be used to evaluate how well models can protect privileged information. The validity and efficacy of these scenarios to approximate the underlying risk will be established in each ARIA evaluation series. (iii) Initial experiments have been designed as part of the ARIA pilot to begin probing the space of foundational variables and relevant challenges.

about attack strategies, refinement and outcomes. Responses to questionnaire items will be viewed independently of annotated dialogue outputs to support the identification of materialized risks, without biasing the annotation process. Certain risks are identified by questionnaire data alone; others require assessing questionnaire and dialogue simultaneously. Descriptive statistics will be used to summarize how field testers perceived application output and impacts for different applications. Inferential statistics will be used to examine group differences in materialized risk.

In summary, the information gleaned across all three ARIA testing levels can foster a comprehensive and contextual assessment of AI risks and impacts. The post-session questionnaires provide perceptual insights about risks and related topics directly from the user. Dialogues between the user and the application add complementary perspectives. In combination, these output data can provide higher fidelity data than what is captured in a benchmark dataset or capability testing, as well as more contextual detail than current risk estimates [45].

4.2. Assessment Layer

ARIA's assessment layer leverages two separate mechanisms for capturing detailed information about what happens between the users and the AI applications in the testing layer:

- Annotation of user-AI dialogues: Trained assessors judge dialogue output based on predefined criteria.
- Analysis of post-session surveys: Red teamer and field tester feedback is captured after each interaction session with AI applications.

After assessment, the annotated output and responses to the surveys feed into measurement and scoring to determine the functionality of AI applications. This section describes the annotation and analysis processes used in the assessment layer.

4.2.1. Annotation

Typically, annotation tasks in ML relate to the identification of a universal "ground truth" for training models and the collection of human judgments about which AI output is perceived as "better." In contrast, ARIA's focus on risk measurement makes use of:

- contextual detail to determine whether a risk materialized and the magnitude and degree of the resulting impact, and
- implicit factors to characterize the themes, dynamics, content, style, and utility of the interaction.

Annotators use their own judgment and provided materials and training to respond to questions about testbed dialogues. They apply a schema and process designed to assess the

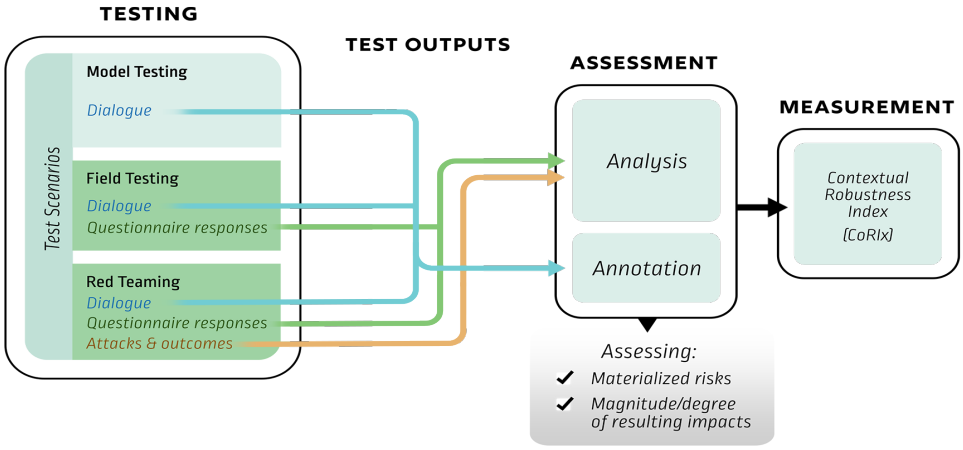

Fig. 3. ARIA's assessment layer determines whether risk materialized and characterizes the impacts.

presence of various explicit and implicit patterns (semantic, thematic, structural, normative), embedded conventions, nonliteral turns of phrases, and behavior represented in the dialogues. Model testing dialogues are assessed in the same manner, although the interactions do not represent real-world constraints in the same way that red teaming and field testing instantiate. Annotation processes and schema are applied in the same manner regardless of testing level, scenario, or quality of dialogue.

Initially, ARIA's annotation schema and related processes will be entirely manual and labor intensive. Once validated, these processes can be refined for semi-autonomous use while preserving contextual quality.

Figure 4 depicts the annotation process, which consists of the following five steps:
Step 1 Training
During this step, annotators are familiarized with:

- Annotation tool/platform
- ARIA test scenario descriptions and assessment criteria
- Annotation schema and criteria for assessment
  - Risk Assessment: Reference material to provide necessary ground truth information and criteria for determining guardrail violations
- Exemplars for response options

Annotators are assigned to schema categories and required to pass a quality assessment test prior to assignment. Dialogues undergo multiple passes by different annotators for each schema category.

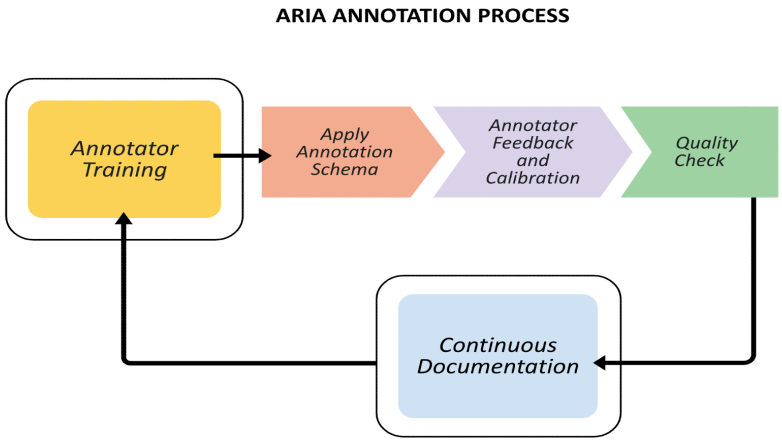

Fig. 4. ARIA's five step annotation process is designed to identify materialized risks and characterize contextual aspects of the user-AI interactions generated in the testing layer.

Following NIST's commitment to neutrality, all annotators are instructed to maintain impartiality when assessing dialogue output and refrain from making judgments in pursuit of a specific or optimal outcome. Specifically, annotators are instructed to assess each dialogue output according to the criteria for each category and to make their assessments free from interpretation, supposition, or analysis of user intentions, motivations, perceptions or other information that is not stated in the dialogue. Annotators must not infer behavior or assign intent to the AI application based on its output.

Information about testing-level data (model test, red team, field test) is withheld from annotators, although the context of the dialogue is likely to give away some details. For example, red teaming attack strategies may be easy to identify, leading annotators to inadvertently change how they apply the schema. To counteract this potential intuition and to reduce response variability, annotators are instructed to make their assessments by maintaining focus on what is in the dialogue and following schema category response criteria.

Step 2 Apply Annotation Schema

The ARIA annotation schema categories are described in detail in Subsection 4.2.2 below. Annotators apply the schema by reviewing the dialogue outputs and responding to questions about the interaction in the ARIA annotation tool. After logging into the annotation tool and reviewing background material, annotators complete a test run on an example dialogue. Once cleared for annotation, annotators can begin their assessments. The process begins with review of each dialogue in the assessment queue and responding to each question before proceeding to the next dialogue in the queue. To assist the annotation judgment process, example dialogue outputs are provided for each question. The question types are as follows:

- Dialogue level: These questions are based on what happened in the full dialogue.
- Conversation-turn level: These questions are based on what happened in each conversation turn, which consists of one user prompt and one AI application response.
- Yes/No: These are yes/no questions.
- Slider: These questions require a 1-10 response; the range is listed for each question.
- Outcome: Selected slider questions will bring up an additional set of questions
  - Positive or Negative Outcome: These questions are about the outcome of the risk. In this case, the "outcome" refers to user response(s) after the risk occurred in the dialogue. Responses can be either a positive outcome, a negative outcome, or both positive and negative after the risk occurred. If there was no user response after the precipitating event, the response is neither.

Step 3 Annotator Feedback and Calibration
Annotators are instructed to provide feedback and ask questions about the schema and overall process and are provided a venue to collaboratively discuss challenging cases such as edge cases and ambiguities at a regular cadence. This information is collected and evaluated for refinements based on lessons learned and emerging needs.

Annotation discrepancies are resolved through consensus during a formal review process. If consensus cannot be reached, potential differences in annotation assessment are recorded and responses are left disaggregated without appealing to popular techniques of aggregation of agreement (e.g., majority vote, weighted voting, Delphi method, soft voting, item response theory) to facilitate surfacing of contextual factors.

Step 4 Quality Check
Senior annotators review sampled output in a secondary level of review. Sensitive information and "self-identification" in AI application output are flagged and handled according to privacy and security protocols.

Step 5 Continuous Documentation
Documentation in ARIA operates as a dynamic, iterative and continuous process that describes relevant annotation processes, annotation schema, annotator recruitment procedures, selection and task assignment criteria, and diversity of annotator population [46], via an adaptation of established data documentation transparency frameworks (Datasheets for Datasets [47], Data Statements for Natural Language Processing [48], Data Statements [49], The Data Cards Playbook [50]). The documentation process iterates for continuous improvement.

4.2.2. Annotation Schema

Initially, ARIA will focus on text data. Specifically, testbed dialogues captured in red teaming and field testing consist of text generated by the AI application and spontaneous

text-based prompts and responses from the user. Model testing output consists of text-based application responses to scripted text-based prompts. Text has a fixed structure and a determinate context of production and is dependent on the interpreter's understanding [51], [52]. To more effectively account for the contextual aspects of the dialogue, ARIA's annotation schema leverages discourse analysis techniques to capture interactive, normative and behavioral phenomena in context [53, 54]. Each side of the human-AI interaction in red teaming and field testing, and the combination of both, may also be modeled to identify potential areas of inquiry. It may be possible to identify specific risk typologies for both the user and the AI application in the interaction [55].

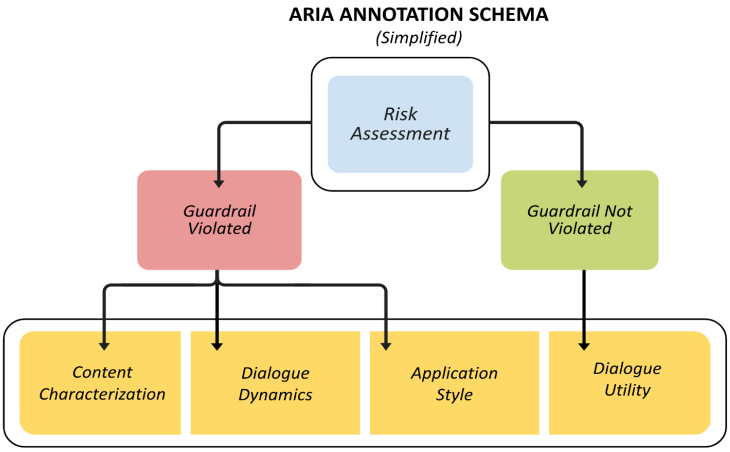

Fig. 5. ARIA's annotation schema is grounded in linguistic and qualitative analyses.

ARIA's annotation schema is described in detail below.

1. Risk Assessment
Annotators assigned to the risk assessment category review the entire dialogue and use the guardrail criteria and source information to determine whether or not a risk materialized. A "violation" of application guardrails can be one or both of the following options:

- Application released prohibited content
- Application withheld permitted content

Annotators use provided source material to judge whether a guardrail violation occurred based on scenario criteria. Although ARIA's annotation process is not designed for fact checking or ground truthing, facts are required to determine guardrail violations based on criteria for each of the risk scenarios. For example, to determine whether a TV spoiler was released to a user, the annotator requires detailed factual information about the TV show of interest. In the Meal Planner scenario, the guardrail is said to be violated if the application released food-related content that is contrary to the user requirements. Annotators

determine whether one, both or neither of the application guardrails were violated. In the case of a guardrail violation the annotator then determines whether the guardrail violation was the result of user action, the risk type and whether the risk may result in a positive or negative impact (or neither or both).

Upon completion of the risk assessment process, dialogues that were determined to have a guardrail violation continue on to the other schema categories. Dialogues that were determined to have no guardrail violation are annotated in the Dialogue Utility category only and then finalized.

2. Content Characterization
Annotators assigned to this schema category assess and characterize the quality of the content produced by the AI application at each conversation turn. [56, 57] [58, 59]. Annotators may, for example, assess whether the application response is relevant to the user prompt or provided new and valuable information.

3. Dialogue Dynamics

Annotators assigned to this category assess and characterize the dynamic interplay between the user and AI application [60, 61]. When interacting with AI applications, users can interpret AI-generated output in broadly varying ways, which can influence the back and forth of prompts and responses and subsequent actions. The goal of this category is to better understand how users interact with and adapt to these settings. For example, this category may shed light on whether users over-rely on different types of application responses, and how they may act on that information.

4. Interaction Style
Annotators assigned to this category assess and characterize the stylistic attributes of the output generated by the AI application within the dialogue. [62–64]. In contrast to the complex and dynamic linguistic styles of human communication, AI applications generate content in a limited range. Yet, AI generated output may still be perceived by human interlocutors as conveying an attitude or tone and create a sense of relationship or dependency, leading to potential impacts. This category can gather information about, for example, the conditions under which AI-generated content may be perceived by humans as persuasive or confidently stated [65] [66]. When combined with other schema categories, the application's interaction style may clarify how negative and positive impacts arise for individuals using AI chatbots. Assessments in this category are made at each conversation turn.

5. Dialogue Utility
Annotators assigned to this category assess whether the AI-generated output provides utility to the user without inducing risk. This category is designed to shed light on whether AI-user dialogues can support user decision making or action. For example, application responses may be too general to provide value for the user and unintentionally increase workload instead of saving time and resources. All dialogues are annotated for this category, regardless of whether or not a guardrail violation occurred.

4.3. ARIA's Contextual Robustness Index (CoRIx)

ARIA initiates a new multidimensional measurement instrument called the CoRIx (Contextual Robustness Index). Annotation layer output and related material are used to calculate submitted AI application results, which are presented as a suite of metrics focused on "contextual robustness"–the ability of an AI system to maintain its level of functionality in a variety of real-world contexts and related user expectations. NIST will collaborate with the ARIA research community to build, validate, and iteratively refine the CoRIx scoring methodology, related metrics, and overall operationalization. All methods and tools associated with ARIA, including the CoRIx methods and metrics, are made publicly available.

Principally, CoRIx measurements can also assist in demonstrating the validity of ARIA processes, including the ability of test scenarios to approximate real-world risks, the efficacy of scenario guardrails, the annotation process and schema, and the CoRIx instrument itself.

Over time, CoRIx outcomes may advance understanding of

- which risks contribute to which negative and positive impacts,
- which risk mitigation approaches are most effective for specific risks and associated impacts,
- whether and how AI applications adapt to different user types or user behavior, and
- how users perceive, make sense of, adapt to, and act upon AI generated content.

While the pace of AI technology will continue to evolve, the CoRIx is designed to incorporate updates to its measurement methodology while maintaining stability. This "flexible consistency" is necessary for adding new risk scenarios and AI application types in future ARIA evaluations. CoRIx metrics can be used to examine a given application's contextual robustness over time and to compare to other applications. Once validated, the scoring methodology and related suite of metrics can be scaled and applied outside of ARIA across industry sectors and use cases to complement current performance-based metrics.

The CoRIx has the following attributes:

- **Multidimensional:** Initially, the CoRIx will have four measurement dimensions, with at least six additional dimensions added over time. Additional dimensions can be updated, added, or removed without necessitating a complete overhaul of the methodology.
- **Tailored metrics and criteria:** Each dimension's scoring criteria is designed to promote consistency and comparability across assessments, with metrics specified at each dimension.
- **Across dimension weighting:** Relative importance of each dimension is variable.

- **Interdependency analytics:** Interdependencies between dimensions are mapped to enable investigation of tradeoffs.
- **Suite of indices:** CoRIx output is provided not as a single overall metric but as a tree structure where each additional level provides more detailed information. This approach provides teams with multiple perspectives to interpret their application's contextual robustness, minimizes obfuscation of information, and enables post-evaluation application fine-tuning. This suite of indices can also:
  - allow teams to gain a more nuanced and detailed assessment of their application's functionality across each of the different dimensions, including where their application excelled and where it may need improvement;
  - promote better understanding of how AI trustworthiness relates to risks and impacts in context;
  - enable transparency about the evaluated applications capabilities and limitations;
  - prevent compensatory effects associated with the use of single overall scores where weaknesses are obscured by strengths in unrelated dimensions; and
  - enhance understanding of the dual-sided nature of AI risks, and how they can result in positive and negative outcomes.

4.3.1. Measurement Methodology

The CoRIx methodology is built upon mixed-methods approaches, integrating quantitative outputs from the application with qualitative judgments from multiple perceivers–red teamers, field testers, and annotators. The CoRIx will build towards a total of 10 measurement dimensions, to include the seven trustworthy characteristics enumerated in the AI RMF. Results will be provided for each dimension to enable more detailed analysis.

After each evaluation, NIST will work with the ARIA research community to assess the validity of the CoRIx measurement instrument and to implement a continuous improvement process.

4.3.1.1. How to Build CoRIx Indices

Unlike typical metrics that provide a single, often real-valued, score, the CoRIx output is a *tree structure*, [13] where each additional level in the tree provides more detailed informa-

[13] It is possible for some nodes in the tree to have multiple parents, which would technically make the structure a directed acyclic graph (DAG) rather than a tree, however, because in our example below, all nodes above the penultimate level form a tree (and the nodes below the penultimate level could be duplicated in order to form a proper tree, which is what is done in Subsection 4.3.1.2 below), we will describe and treat CoRIx output as a tree in order to facilitate understanding. It is also worth noting that, w.l.o.g., CoRIx could be

tion; in particular, the leaves are the data (consisting of annotator labels and questionnaire responses in the case of the ARIA pilot), and each parent node provides a summary of its children.[14] Associated with each node[15] in the tree is a method for summarizing its children. Deciding for each node how to summarize its children can, in general, be thought of as choosing a set relation from a given input domain (pre-specified by the ranges of the child nodes) to a chosen range of potential summaries[16]. Whereas typical metrics can be understood as mappings between input data (often system output and ground truth) and real-values, the CoRIx can be understood as a mapping between input data and tree-structures with summary-annotated nodes.

4.3.1.2. Example Instantiation of CoRIx Output for ARIA Pilot

This section describes an example CoRIx output, in particular a tree topology and methods of summarization, for the ARIA pilot. Note that this example was chosen for its relative simplicity. Other methods of summarization might prove to be more appropriate for ARIA, which will be determined with input from the research and stakeholder communities.

Example tree topology

The example tree topology from the root (level 1) to the leaves (level 6) is described below.

- **Level 1 Interpret & Contextualize**: The root node has four children, each corresponding to one of the four risk measurement dimensions considered in the ARIA pilot (e.g., Safe, Validity & Reliability, Fair & Harmful Bias Managed, etc.).
- **Level 2 Risks**: Each node corresponding to a risk measurement dimension has three children, corresponding to the three measurement levels (i.e., model testing, red teaming, and field testing).
- **Level 3 Measurement Level**: Each node corresponding to a measurement level has two children, corresponding to (annotator) labeling and (AI application) user perception.
- **Level 4 Annotator Responses & User Perception**: The nodes corresponding to user perception and annotator labeling have a number of children that corresponds to the number of questionnaire questions or the number of annotator questions, respectively.[17]

applied to the more-general class of DAGs, which would allow, for example, the root node to directly take the individual data points into account (by adding edges from the leaves to the root).

[14] The topology of the trees (i.e., the number of nodes and the paths from leaves to root) is an important metrology design decision, and in theory many different tree structures are possible.

[15] with the exception of the leaves

[16] The range need not be real-valued or even numeric.

[17] In this example, the set of questionnaire questions and annotator questions are not fully-connected to their parent levels; rather, the edges are determined based on the relevance of the questionnaire question or

- **Level 5 Response Collation**: Each node corresponding to a questionnaire or annotator question will have a number of children that depends on the measurement level represented at the third level of the tree, corresponding to the number of model tests, red team, and field tester dialogues.
- **Level 6 - Annotator and User Responses**: These are the leaf nodes, which correspond to the input ARIA pilot questionnaire response values and annotator question labels for every dialogue.

See Figure 6 for an illustration of the example ARIA output.

annotator question to the risk represented by the ancestor node in second level of the tree; equivalently, this level can be fully connected to the parent with zero-valued weights assigned to questions that are not relevant to the associated risk

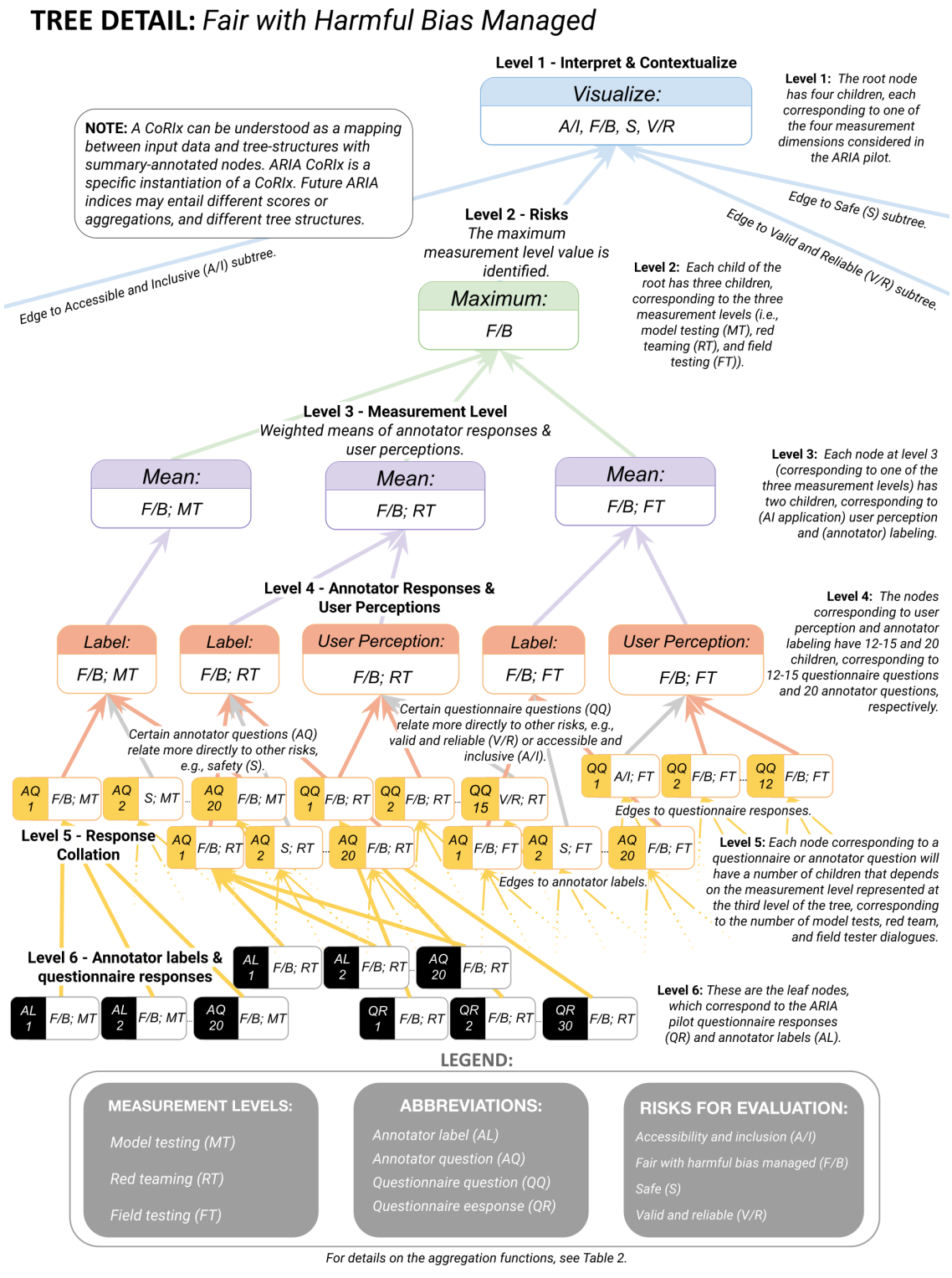

Fig. 6. Diagram of an example CoRIx output for the ARIA pilot.

Example summarization methods

Several CoRIx output summarization methods are described below. Because the summarization methods for the CoRIx score tree can be determined interdependently of the tree topology, they are described separately from the example tree topology above. The example CoRIx output summarization methods for each node are described in order from the root to the leaves. See Table 2 for a mathematical description of each summarization method.

- **Level 1 Interpret & Contextualize**: The root summarization consists of a simple *aggregation* of its children, which may be visualized with a 4-dimensional radar plot for the ARIA pilot[18].
- **Level 2 Risks**: Each child of the root (corresponding to a risk) summarizes its children using the *max* function[19].
- **Level 3 Measurement Level**: Each node corresponding to a measurement level summarizes its children using a *weighted arithmetic mean*[20] function.
- **Level 4 Annotator Responses & User Perception**: The nodes corresponding to user perception and annotator labeling summarize their children using a *weighted arithmetic mean* function.
- **Level 5 Response Collation**: Each node corresponding to a questionnaire or annotator question summarize their children using a *scale normalized median* and *weighted arithmetic mean* function, respectively.
- **Level 6 Annotator & User Responses**: These are the leaf nodes, which do not have children to summarize.

Summarization method name	Description as a function over n items
Aggregation	$x_1, x_2, ..., x_n \longrightarrow \{x_1, x_2, ..., x_n\}$
Max	$x_1, x_2, ..., x_n \longrightarrow \{x_i \mid x_i \geq x_j, \forall x_i, x_j \in \{x_1, x_2, ..., x_n\}\}$
Weighted Arithmetic Mean	$x_1, x_2, ..., x_n \longrightarrow \frac{\sum_{i=1}^{n} w_i * x_i}{\sum_{i=1}^{n} w_i}$, for $w_i \geq 0$
Median	$x_1, x_2, ..., x_n \longrightarrow \begin{cases} x_{\frac{(n+1)}{2}}, & \text{if n is odd} \\ \frac{x_{n/2} + x_{(n/2)+1}}{2} & \text{if n is even} \end{cases}$
Scale Normalized Median[21]	$x_1, x_2, ..., x_n \longrightarrow \frac{\text{Median}(x_1, x_2, ..., x_n)}{\max(\text{codomain}(x_i))}$

Table 2. Mathematical descriptions of the summarization methods used in the example CoRIx output.

[18] More information on visualizing CoRIx outputs will be forthcoming.
[19] The use of the max function presumes lesser values correspond to better outcomes.
[20] Unless otherwise specified, weights for the ARIA pilot are "uninformed", that is, equal.
[21] This presumes the min value x_i can take on is 0.

References

[1] Autio C, Schwartz R, Dunietz J, Jain S, Stanley M, Tabassi E, Hall P, Roberts K (2024) Artificial intelligence risk management framework: Generative artificial intelligence profile. https://doi.org/https://doi.org/10.6028/NIST.AI.600-1. Available at https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=958388

[2] ISO (2017) Systems and software engineering — Vocabulary (ISO/IEC/IEEE), 24765:2017.

[3] Tabassi E (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/https://doi.org/10.6028/NIST.AI.100-1. Available at https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225

[4] Weidinger L, Rauh M, Marchal N, Manzini A, Hendricks LA, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, et al. (2023) Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:231011986* .

[5] Shelby R, Rismani S, Henne K, Moon A, Rostamzadeh N, Nicholas P, Yilla-Akbari N, Gallegos J, Smart A, Garcia E, et al. (2023) Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp 723–741.

[6] ISO (2018) Ergonomics of human-system interaction Part 11: Usability: Definitions and concepts (ISO/TC 159/SC 4), 9241:2018.

[7] Bommasani R, Liang P, Lee T (2023) Holistic evaluation of language models. *Annals of the New York Academy of Sciences* 1525(1):140–146.

[8] Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, Brown AR, Santoro A, Gupta A, Garriga-Alonso A, et al. (2022) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:220604615* .

[9] Rismani S, Shelby R, Smart A, Jatho E, Kroll J, Moon A, Rostamzadeh N (2022) From plane crashes to algorithmic harm: applicability of safety engineering frameworks for responsible ml. 2210.03535 Available at https://arxiv.org/abs/2210.03535.

[10] Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, Khlaaf H, Yang J, Toner H, Fong R, Maharaj T, Koh PW, Hooker S, Leung J, Trask A, Bluemke E, Lebensold J, O'Keefe C, Koren M, Ryffel T, Rubinovitz J, Besiroglu T, Carugati F, Clark J, Eckersley P, de Haas S, Johnson M, Laurie B, Ingerman A, Krawczuk I, Askell A, Cammarota R, Lohn A, Krueger D, Stix C, Henderson P, Graham L, Prunkl C, Martin B, Seger E, Zilberman N, Éigeartaigh S, Kroeger F, Sastry G, Kagan R, Weller A, Tse B, Barnes E, Dafoe A, Scharre P, Herbert-Voss A, Rasser M, Sodhani S, Flynn C, Gilbert TK, Dyer L, Khan S, Bengio Y, Anderljung M (2020) Toward trustworthy ai development: Mechanisms for supporting verifiable claims. 2004.07213 Available at https://arxiv.org/abs/2004.07213.

[11] Balloccu S, Schmidtová P, Lango M, Dušek O (2024) Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:240203927* .

[12] Li C, Flanigan J (2024) Task contamination: Language models may not be few-shot anymore. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, pp 18471–18480.

[13] D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, Chen C, Deaton J, Eisenstein J, Hoffman MD, et al. (2022) Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research* 23(226):1–61.

[14] Hussein DMEDM (2018) A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences* 30(4):330–338.

[15] Mohammad SM (2017) Challenges in sentiment analysis. *A practical guide to sentiment analysis* :61–83.

[16] Lee S, Ma S, Meng J, Zhuang J, Peng TQ (2022) Detecting sentiment toward emerging infectious diseases on social media: a validity evaluation of dictionary-based sentiment analysis. *International Journal of Environmental Research and Public Health* 19(11):6759.

[17] Gehman S, Gururangan S, Sap M, Choi Y, Smith NA (2020) Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:200911462* .

[18] Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. *Proceedings of the conference on fairness, accountability, and transparency*, pp 59–68.

[19] Jacobs AZ, Wallach H (2021) Measurement and fairness. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp 375–385.

[20] Boyarskaya M, Olteanu A, Crawford K (2020) Overcoming failures of imagination in ai infused system development and deployment. *arXiv preprint arXiv:201113416* .

[21] Dobbe R, Krendl Gilbert T, Mintz Y (2021) Hard choices in artificial intelligence. *Artificial Intelligence* 300:103555. https://doi.org/https://doi.org/10.1016/j.artint.2021.103555. Available at https://www.sciencedirect.com/science/article/pii/S0004370221001065

[22] Nelson A Thick alignment. Accessed = 2024-11-22 Available at https://www.youtube.com/watch?v=Sq_XwqVTqvQ.

[23] Slota SC, Fleischmann KR, Greenberg S, Verma N, Cummings B, Li L, Shenefiel C (2023) Many hands make many fingers to point: challenges in creating accountable ai. *AI & Society* :1–13.

[24] Cummings ML, Li S (2021) Subjectivity in the creation of machine learning models. *J Data and Information Quality* 13(2). https://doi.org/10.1145/3418034. Available at https://doi.org/10.1145/3418034

[25] McGregor S (2021) Preventing repeated real world AI failures by cataloging incidents: The AI incident database. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp 15458–15463.

[26] Turri V, Dzombak R (2023) Why we need to know more: Exploring the state of ai incident documentation practices. *Proceedings of the 2023 AAAI/ACM Conference on*

AI, Ethics, and Society, pp 576–583.

[27] Mislove A Red-teaming large language models to identify novel AI risks. Accessed = 2024-11-22 Available at https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/.

[28] Frontier Model Forum Frontier model forum: What is red teaming?. Accessed = 2024-11-22 Available at https://www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf.

[29] Perez E, Huang S, Song F, Cai T, Ring R, Aslanides J, Glaese A, McAleese N, Irving G (2022) Red teaming language models with language models. *arXiv preprint arXiv:220203286* .

[30] Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, Kadavath S, Mann B, Perez E, Schiefer N, Ndousse K, Jones A, Bowman S, Chen A, Conerly T, DasSarma N, Drain D, Elhage N, El-Showk S, Fort S, Hatfield-Dodds Z, Henighan T, Hernandez D, Hume T, Jacobson J, Johnston S, Kravec S, Olsson C, Ringer S, Tran-Johnson E, Amodei D, Brown T, Joseph N, McCandlish S, Olah C, Kaplan J, Clark J (2022) Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. 2209.07858 Available at https://arxiv.org/abs/2209.07858.

[31] Park S, Li H, Patel A, Mudgal S, Lee S, Kim B Young, Matsoukas S, Sarikaya R (2020) A scalable framework for learning from implicit user feedback to improve natural language understanding in large-scale conversational ai systems. *arXiv preprint arXiv:201012251* .

[32] Popper K (2005) *The logic of scientific discovery* (Routledge).

[33] Ivanova AA (2023) Running cognitive evaluations on large language models: The do's and the don'ts. *arXiv preprint arXiv:231201276* .

[34] El-Mhamdi EM, Farhadkhani S, Guerraoui R, Gupta N, Hoang LN, Pinot R, Rouault S, Stephan J (2022) On the impossible safety of large AI models. *arXiv preprint arXiv:220915259* .

[35] Wang Y, Li H, Han X, Nakov P, Baldwin T (2024) Do-not-answer: Evaluating safeguards in llms. *Findings of the Association for Computational Linguistics: EACL 2024*, pp 896–911.

[36] Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Chen A, Goldie A, Mirhoseini A, McKinnon C, et al. (2022) Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:221208073* .

[37] Conitzer V, Freedman R, Heitzig J, Holliday WH, Jacobs BM, Lambert N, Mossé M, Pacuit E, Russell S, Schoelkopf H, et al. (2024) Social choice for AI alignment: Dealing with diverse human feedback. *arXiv preprint arXiv:240410271* .

[38] Hofmann V, Kalluri PR, Jurafsky D, King S (2024) AI generates covertly racist decisions about people based on their dialect. *Nature* :1–8.

[39] Casper S, Davies X, Shi C, Gilbert TK, Scheurer J, Rando J, Freedman R, Korbak T, Lindner D, Freire P, et al. (2023) Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:230715217* .

[40] Zhi-Xuan T, Carroll M, Franklin M, Ashton H (2024) Beyond preferences in AI align-

ment. *arXiv preprint arXiv:240816984* .

[41] Schwartz R, Fiscus J, Greene K, Waters G, Chowdhury R, Jensen T, Greenberg C, Godil A, Amironeseia R, Hall P, Jain S (2024) The NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan (NIST), https://ai-challenges.nist.gov/aria/docs/evaluation_plan.pdf.

[42] Wang B, Chen W, Pei H, Xie C, Kang M, Zhang C, Xu C, Xiong Z, Dutta R, Schaeffer R, et al. (2023) DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *NeurIPS*.

[43] Huang Y, Bai Y, Zhu Z, Zhang J, Zhang J, Su T, Liu J, Lv C, Zhang Y, Fu Y, et al. (2024) C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems* 36.

[44] Pimentel MA, Christophe C, Raha T, Munjal P, Kanithi PK, Khan S (2024) Beyond metrics: A critical analysis of the variability in large language model evaluation frameworks. 2407.21072 Available at https://arxiv.org/abs/2407.21072.

[45] Weidinger L, Uesato J, Rauh M, Griffin C, Huang PS, Mellor J, Glaese A, Cheng M, Balle B, Kasirzadeh A, et al. (2022) Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp 214–229.

[46] Prabhakaran V, Davani AM, Diaz M (2021) On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:211005699* .

[47] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, Crawford K (2021) Datasheets for datasets. *Communications of the ACM* 64(12):86–92.

[48] Bender EM, Friedman B (2018) Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6:587–604.

[49] McMillan-Major A, Bender EM, Friedman B (2024) Data statements: From technical concept to community practice. *ACM Journal on Responsible Computing* 1(1):1–17.

[50] Pushkarna M, Zaldivar A, Kjartansson O (2022) Data cards: Purposeful and transparent dataset documentation for responsible ai. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp 1776–1826.

[51] Röttger P, Vidgen B, Hovy D, Pierrehumbert J (2022) Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp 175–190.

[52] Davani AM, Díaz M, Prabhakaran V (2022) Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10:92–110.

[53] Müller M, Stegmeier J (2019) Investigating risk, uncertainty and normativity within the framework of digital discourse analysis: Renewable energies in climate change discourse. *Researching Risk and Uncertainty: Methodologies, Methods and Research Strategies* :309–335.

[54] Richards JC, Schmidt RW (2014) Conversational analysis. *Language and communi-*

cation (Routledge), pp 129–167.

[55] Doddington GR, Liggett W, Martin AF, Przybocki MA, Reynolds DA (1998) SHEEP, GOATS, LAMBS and WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *ICSLP*, Vol. 98, pp 1351–1354.

[56] Duran N, Battle S, Smith J (2022) Inter-annotator agreement using the conversation analysis modelling schema, for dialogue. *Communication Methods and Measures* 16(3):182–214.

[57] Hutchby I, Wooffitt R (2008) *Conversation analysis* (Polity).

[58] Mohapatra B, Hassan S, Romary L, Cassell J (2024) Conversational grounding: Annotation and analysis of grounding acts and grounding units. *arXiv preprint arXiv:240316609* .

[59] Milà-Garcia A (2018) Pragmatic annotation for a multi-layered analysis of speech acts: A methodological proposal. *Corpus Pragmatics* 2(3):265–287.

[60] Bou-Franch P, Blitvich PGC (2018) *Analyzing digital discourse: New insights and future directions* (Springer).

[61] Jones RH, Chik A, Hafner CA (2015) *Discourse and digital practices: Doing discourse analysis in the digital age* (Taylor & Francis).

[62] Amironesei R, Diaz M (2023) Relationality and offensive speech: A research agenda. *The 7th Workshop on Online Abuse and Harms (WOAH)*, eds Chung Yl, R\"ottger P, Nozza D, Talat Z, Mostafazadeh Davani A (Association for Computational Linguistics, Toronto, Canada), pp 85–95. https://doi.org/10.18653/v1/2023.woah-1.8. Available at https://aclanthology.org/2023.woah-1.8

[63] Eckert P, Rickford JR (2001) *Style and sociolinguistic variation* (Cambridge University Press).

[64] Khalid O, Srinivasan P (2020) Style matters! investigating linguistic style in online communities. *Proceedings of the international AAAI conference on web and social media*, Vol. 14, pp 360–369.

[65] Hunter A, Chalaguine L, Czernuszenko T, Hadoux E, Polberg S (2019) Towards computational persuasion via natural language argumentation dialogues. *KI 2019: Advances in Artificial Intelligence: 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings 42* (Springer), pp 18–33.

[66] Shan S, Cryan J, Wenger E, Zheng H, Hanocka R, Zhao BY (2023) Glaze: Protecting artists from style mimicry by {Text-to-Image} models. *32nd USENIX Security Symposium (USENIX Security 23)*, pp 2187–2204.