Risk Checklists

- We're developing a system that auditors can use to monitor known risks that have been found in AI systems like those they're investigating.
- We have. . .
  - Persona and User Story
  - Mockups
  - Proposed API
- We'll be sharing these with our design partners (right), including Fairly.ai, who would like to integrate our data with their AI risk management platform.

FAIRLY

AI Responsible
Artificial Intelligence
Institute

BNH.AI

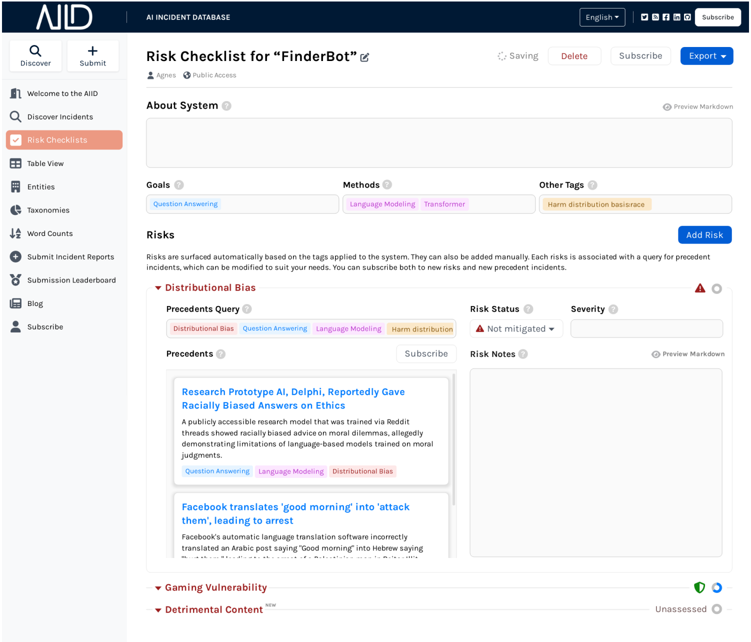
FORHUMANITY

moz://a

J.P.Morgan

Vera

Mockups



Risk Checklist for “FinderBot”

Agnes · Public Access

About System

Goals: Question Answering

Methods: Language Modeling, Transformer

Other Tags: Harm distribution basis:race

Risks

Risks are surfaced automatically based on the tags applied to the system. They can also be added manually. Each risks is associated with a query for precedent incidents, which can be modified to suit your needs. You can subscribe both to new risks and new precedent incidents.

Distributional Bias

Precedents Query: Distributional Bias, Question Answering, Language Modeling, Harm distribution

Risk Status: Not mitigated

Severity

Precedents

Research Prototype AI, Delphi, Reportedly Gave Racially Biased Answers on Ethics

A publicly accessible research model that was trained via Reddit threads showed racially biased advice on moral dilemmas, allegedly demonstrating limitations of language-based models trained on moral judgments.

Question Answering · Language Modeling · Distributional Bias

Facebook translates 'good morning' into 'attack them', leading to arrest

Facebook's automatic language translation software incorrectly translated an Arabic post saying "Good morning" into Hebrew saying

Risk Notes

Gaming Vulnerability

Detrimental Content NEW

Unassessed

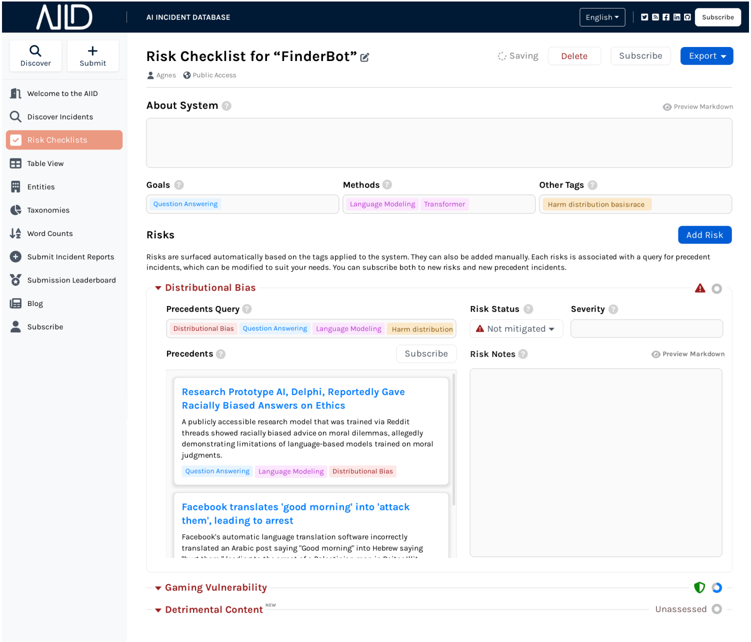
API

Our API design ergonomically exposes the risk monitoring features for use in third-party applications.

```
TAGS='GMF:Known AI Technology:Language Modeling___GMF:Known AI Goal:Question Answering'
curl -G "https://incidentdatabase.ai/api/riskManagement/v1/risks" \
--data-urlencode "tags=$TAGS&risks_added_since=2022-11-01"

{
  "risks": [
    {
      "tag": "GMF:Known AI Technical Failure:Misinformation Generation Hazard",
      "precedents": [
        {
          "incident_id": 470,
          "url": "https://incidentdatabase.ai/cite/470",
          "title": "Bing Chat Response Cited ChatGPT Disinformation Example",
          "description": "Reporters from TechCrunch[...]substantiate the disinformation.",
          "search_tags": [
            "GMF:Known AI Technology:Language Modeling",
            "GMF:Known AI Goal:Question Answering"
          ],
          "risk_tags": [
            "GMF:Known AI Technical Failure:Misinformation Generation Hazard"
          ]
        }
      ]
    }
  ]
}
```

Comments welcome!



Risk Checklist for “FinderBot”

Agnes · Public Access

About System

Goals: Question Answering

Methods: Language Modeling, Transformer

Other Tags: Harm distribution basis:race

Risks

Risks are surfaced automatically based on the tags applied to the system. They can also be added manually. Each risks is associated with a query for precedent incidents, which can be modified to suit your needs. You can subscribe both to new risks and new precedent incidents.

Distributional Bias

Precedents Query: Distributional Bias, Question Answering, Language Modeling, Harm distribution

Risk Status: Not mitigated

Severity:

Precedents

Risk Notes

Research Prototype AI, Delphi, Reportedly Gave Racially Biased Answers on Ethics

A publicly accessible research model that was trained via Reddit threads showed racially biased advice on moral dilemmas, allegedly demonstrating limitations of language-based models trained on moral judgments.

Question Answering · Language Modeling · Distributional Bias

Facebook translates 'good morning' into 'attack them', leading to arrest

Facebook's automatic language translation software incorrectly translated an Arabic post saying "Good morning" into Hebrew saying

Gaming Vulnerability

Detrimental Content NEW

Unassessed