Impact Assessment / Harms Analysis Templates

AUGUST 23, 2023
CDAO PWP 6-761A5 Task 2.3
AI Division, Research and Development, AI/ML
Introductions & PWP Overview
SEI AI Division (initial) Team

Principal Investigator
Carol J. Smith
Senior Research Scientist
Email: cjsmith@sei.cmu.edu

Responsibilities:
• Guide the team.
• Design and lead appropriate research efforts.
• Contribute expertise to complex technical research tasks.
• Foster collaborative research environment and mentorship.

Design Researcher
Katie Robinson
Assistant Design Researcher
Email: kmrobinson@sei.cmu.edu

Responsibilities:
• Explore user needs for design insights.
• Develop prototypes from refined concepts.
• Collaborate on informed design decisions.
Task 2.3 Overview

Impact Assessment / Harms Analysis Templates

2.3.1. Research & Discovery
SEI will work with the CDAO RAI team and groups across the DoD to identify and study use cases and risk profiles. Analyze information and identify primary use cases and risk profiles to focus on for the effort. Further research and analysis to identify associated needs for each of the selected use cases and risk profiles.

Deliverables
Sept 2023
• Study protocol and materials
• Analysis and mapping of use cases and risk profiles

Oct – Dec 2023
• Primary use cases and risk profiles identified
Task 2.3 Overview

Impact Assessment / Harms Analysis Templates

2.3.2: Artifact Development
Based on the findings from R&D, the SEI will iterate to create appropriate impact assessments and harms analysis artifacts (tools, processes, and templates) for each of the selected use cases and risk profiles. The SEI will create prototypes that are useful and usable to support the selected use cases and risk profiles to enable successful impact assessments and harms analysis efforts.

Deliverables
Oct – Dec 2023
• Initial artifacts for selected use case(s) and risk profile(s)

Jan – March 2024
• Study protocol and materials for 2.3.3
• Additional artifacts for selected use cases and risk profiles
Task 2.3 Overview

Impact Assessment / Harms Analysis Templates

2.3.3 Evaluation and Iteration
Engage with development teams to test and iterate and improve the artifacts through evaluations. The SEI will make improvements to the artifacts and prepare them for transition to CDAO for implementation.

The output of this task will enable more effective and efficient impact assessments and harms analysis efforts for the selected primary use cases and risk profiles.

Deliverables
April – June 2024
• Improved artifacts based on feedback

July – Oct 2024
• Vetted artifacts
• Guidance for development and implementation of artifacts
Discussion
Project Goals

What is the outcome desired from this work?

What impact will the templates have on DoD?

Who will use the templates and how?

What sources of impact or harms should we examine (data, model development, interface/interaction design)?
CDAO Requests

Use cases (DoD and NIST)
Risk profiles (DoD and NIST)
Impact Assessments
Impact Assessment - Assumptions

Include benefits and risks of system (full implementation) to the DoD, to end users, to those affected by the system, and to broader society.

Conducted prior to development (or early stages), and continuously throughout operation.

Different types of systems, contexts, end users, etc. require different types of impact assessments.

Excluding: Very high risk implementations of AI.
Types of Impacts

Positive

• Efficiency / effectiveness
• Better solutions
• User experience / usability

Negative

• Complexity
• Risk
• Harms (next topic)
Existing Impact Assessments

Open Canada: Algorithmic Impact Assessment

IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being

ECP: Artificial Intelligence Impact Assessment (AIIA)

https://open.canada.ca/aia-eia-js/?lang=en


Harms Analysis
What do we mean by harm?

Assumption: Harms are negative.

Types of harms
- Discrimination/prejudice (racism, sexism, etc.)
- Financial
- Privacy
- Physical harm (autonomous vehicles)
- Psychological
- Social harms
- Sustainability/environmental

Excluding: Hardware/Cybersecurity
Microsoft: Types of Harm

- Risk of injury (physical, emotional or psychological injury)
- Denial of consequential services (opportunity loss, economic loss)
- Infringement on human rights (dignity loss, liberty loss, privacy loss, environmental impact)
- Erosion of social & democratic structures (manipulation, social detriment)

### 3Q-Do No Harm Framework: Categories of Harm

<table>
<thead>
<tr>
<th>CATEGORIES OF HARM</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial</td>
<td>Negative impact on finances, property, or other resources</td>
</tr>
<tr>
<td>Health</td>
<td>Negative impact on mental, emotional, or physical health</td>
</tr>
<tr>
<td>Time</td>
<td>Inefficient or unproductive activities, processes, or systems</td>
</tr>
<tr>
<td>Fairness/Equity</td>
<td>Perpetuating or facilitating prejudice, bias and/or unfairness</td>
</tr>
<tr>
<td>Safety</td>
<td>Physical and/or emotional wellbeing compromised by fear, danger, or uncertainty</td>
</tr>
<tr>
<td>Privacy</td>
<td>Lack of control over personal information</td>
</tr>
<tr>
<td>Misinformation</td>
<td>Creation, spread and/or amplification of false or inaccurate information intended to deceive</td>
</tr>
<tr>
<td>Control</td>
<td>Inability to freely direct information, activities, or systems</td>
</tr>
<tr>
<td>Transparency</td>
<td>Lack of disclosure of information, activities, or systems</td>
</tr>
</tbody>
</table>

3Q-Do No Harm Framework, Lisa D. Dance. [https://serviceease.net/3q-do-no-harm-framework](https://serviceease.net/3q-do-no-harm-framework)
Existing Harms Analysis Methods

What Could Go Wrong?
Card Game

3Q-Do No Harm Framework

Microsoft Harms Modeling

What Could Go Wrong?
https://doi.org/10.1145/3409251.3411734

3Q-Do No Harm Framework, Lisa D. Dance.
https://serviceease.net/3q-do-no-harm-framework

Incident Databases

AIID

Where in the World is AI?

AI and algorithmic incident and controversies (AIAAIC)

https://incidentdatabase.ai/

https://map.ai-global.org/

Carol J. Smith, Principal Investigator
Email: cjsmith@sei.cmu.edu

Katie Robinson, Design Researcher
Email: kmrobinson@sei.cmu.edu