Corresponding author: edyta.bogucka@nokia-bell-labs.com
Word Count: 4614 out of 5000 (1000 for 4 figures)

AI Design: A Responsible AI Framework for Pre-filling Impact Assessment Reports

Edyta Bogucka
Nokia Bell Labs, Cambridge (UK)

Marios Constantinides
Nokia Bell Labs, Cambridge (UK)

Sanja Šćepanović
Nokia Bell Labs, Cambridge (UK)

Daniele Quercia
Nokia Bell Labs, Cambridge (UK)

***Abstract*—Impact assessment reports for high-risk AI systems will be legally required but challenging to complete, especially for smaller companies. That is because the current process is complex, costly, and relies on guidebooks with limited assistance. We propose *AI Design*, a semi-automatic framework for pre-filling these reports. It consists of two components: (A) *StakeLinker*, an interactive tool combining various stakeholders' perspectives; and (B) *FillGen*, an LLM-based tool processing stakeholders' perspectives and producing the report to be reviewed by regulatory experts within a company. We conducted two user studies: the first with 13 AI practitioners who confirmed *StakeLinker's* effectiveness in gathering comprehensive input for impact assessment; the second with 8 additional practitioners who successfully evaluated a report for a crime analysis system pre-filled by *FillGen*. To show generalizability, we also made the reports for two other AI systems publicly available.**

Index Terms: K.4.1.c Ethics, K.4.1.d Human safety, I.2 Artificial Intelligence, I.6.5 Model Development, I.6.4 Model Validation and Analysis, O.9.3 Moral implications, O.9.4 Legal implications, O.9.5 Public policy.

INTRODUCTION

As AI governance evolves [20], impact assessment reports will become legally required. The

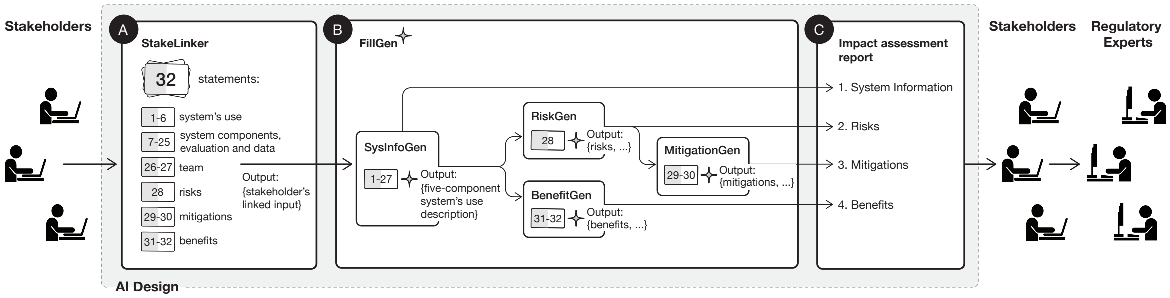


Figure 1: *AI Design* is a framework that guides stakeholders in pre-filling impact assessment reports. Its first component, *StakeLinker* (A), solicits input from multiple stakeholders on various aspects of the AI system, using 32 statements grounded in literature. This input is subsequently fed into *FillGen* (B), an LLM-powered tool. *FillGen* is responsible for summarizing stakeholders' input and producing new insights about the system. Its *SysInfoGen* processes the first 27 statements to generate information about the system's use, while *RiskGen* assesses associated risks, *BenefitGen* identifies potential benefits, and *MitigationGen* suggests mitigations. This content is subsequently used to pre-fill an *impact assessment report* (C). The stakeholders review the pre-filled report, and present it to regulatory experts for final approval.

EU AI Act[1] mandates reports on the impact of high-risk AI systems to enhance transparency on their functionalities and hold companies accountable for ethical and societal impacts. This requirement is really important because, as of now, there have been over 600 actual AI incidents[2] (like an app that made everyone's photos look whiter, and an Amazon feature that was not fair to books by minority authors) and more than 10,000 possible hazards with AI have been cataloged.[3] This shows why we need to carefully check the impact of AI before using it [1] [9].

However, completing impact assessment reports is challenging, especially for smaller companies struggling to keep up with evolving AI regulations.[4] These struggles arise from two factors. First, the assessment process is complicated and expensive. It is complicated because it needs input from different people like developers and lawyers. It is expensive so much so that the certification process will increase development costs by 10-14%[5], adding up to €400 thousand, which is hard for small companies to afford. Second, existing guidebooks that help stakeholders fill in these reports lack practicality as they contain various guiding formats (e.g., FAQs, brainstorming prompts) that might not be relevant to all stakeholders or applicable to a variety of AI uses. Seeing these difficulties, the EU has asked for new tools to help with AI risk assessment[1].

We propose *AI Design*[6], a semi-automatic framework for pre-filling impact assessment reports (Figure 1). It consists of two components: *StakeLinker*, an interactive tool for gathering stakeholders' input about an AI system's use and organizing it for large language model (LLM) processing; and *FillGen*, an LLM-based tool for producing the impact assessment report using *StakeLinker's* inputs. Unlike current solutions that are not designed to work for all AI applications [18] or only cover parts of the assessment [3], *StakeLinker* uses a five-part format that works for any AI application, and *FillGen* handles the entire process of impact assessments, including the identification of risks, mitigations, and benefits. We used *AI Design* to pre-fill a report for an AI system identifying crime hotspots from CCTV footage, demonstrating its potential to simplify impact assessment and reduce compliance burdens. To show its generalizability, we also provided reports for a vocational training

[1] https://artificialintelligenceact.eu
[2] https://incidentdatabase.ai
[3] https://oecd.ai/en/incidents
[4] https://www.digitalsme.eu/the-ai-act-help-or-hindrance-for-smes
[5] https://www.isaca.org/resources/isaca-journal/issues/2023/volume-2/the-potential-impact-of-the-european-commissions-proposed-ai-act-on-smes
[6] Project's website: https://social-dynamics.net/ai-design

chatbot and a system assessing damages after natural disasters.

RELATED WORK

Until recently, there was no consensus on the content of impact assessment reports [16]. Scholars have agreed on the need for three sections (AI system use, risks, and mitigation strategies), and these sections are reflected in four newly proposed templates that we discuss next.

Sherman and Eisenberg introduced the "standardized risk profile" [15], comprised of the system's information, risks, mitigation strategies, formal evaluations, and impact summaries. Similarly, Microsoft's report includes system information, risks, mitigation strategies; and an Algorithmic Impact Assessment from the Ada Lovelace Institute[7] for AI applications in healthcare and NIST's Algorithmic Impact Assessment[8] also cover these aspects, with NIST adding details on system tasks, operational contexts, and organizational risks. However, these reports tend to prioritize risks over benefits. A balanced consideration, though, provides a well-informed perspective, contextualizes risks, and helps prioritize mitigation strategies. Therefore, *AI Design* includes not only system use, risks, and mitigation strategies but also benefits.

A number of guidebooks have been developed to facilitate the completion of these reports. NIST provides comprehensive lists of actions, documents, and glossaries[9], while Microsoft and Ada Lovelace Institute offer group exercises and question prompts. CredoAI provides examples of populated templates but lacks specific guidelines on recreating them [15]. All these works have partly tackled the same problem: when developers start with a blank template and guide, they often experience writer's block because they have to fill in all the content and ensure it meets current regulations. To help with this, earlier work suggested using AI to initially pre-fill in these reports by listing AI users and subjects and providing the necessary legal information [3] [4] [5] [7] [10].

[7]https://www.adalovelaceinstitute.org/resource/aia-template
[8]https://www.equalai.org/aia
[9]https://airc.nist.gov/AI_RMF_Knowledge_Base

AI DESIGN

AI Design is a framework for guiding stakeholders to populate impact assessment reports. Next, we describe its two components— *StakeLinker* and *FillGen*—and how they produce the *impact assessment report*.

A. StakeLinker

This is an interactive tool that facilitates the gathering of input from multiple stakeholders regarding the AI system for which the impact assessment report has to be written. We developed this tool in two steps.

Step 1. Curating a set of statements. To gather stakeholders' information about the AI system's use, we curated 32 statements by reviewing four papers on Responsible AI (RAI) guidelines, questionnaires, and checklists [6] [8] [11] [12], borrowing mostly from [6]. These statements (available on the project's website[6]) systematically collect information about the system's use (statements 1–6), components, evaluation, and data (statements 7–25), involved teams (statements 26–27), risks and mitigations (statements 28–30), and benefits (statements 31–32). For example, one statement asks to identify individuals interacting with the system [6] [12], while another asks to report performance metrics across demographic groups [6] [11] [12].

Step 2. Placing the statements in a newly created UI. To design the interface integrating the 32 statements, we conducted a user study with 13 AI practitioners (2 females, 11 males, median age 39), approved by our institution. Participants work at a large tech company and had 2-10 years of experience in data visualization, data science, machine learning (ML), and natural language processing (NLP).

During the study, participants interacted with two prototype interfaces commonly used in RAI tools: a scroll-page checklist, and a one-page card stack [6] [11]. The checklist presents each statement with a checkbox and an input box (Figure 2A) for users to tick when the statement is implemented and write the implementation details. The stack shows each statement on a two-sided card (Figure 2B), allowing users to mark it as (to be) implemented and note implementation details on the back.

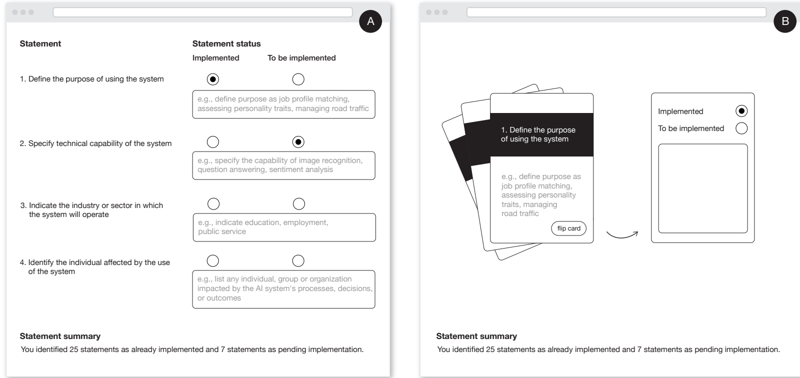


Figure 2: Two preliminary prototypes of *StakeLinker*'s user interface: (A) a scroll-page checklist, and (B) a one-page card stack.

We evaluated the two prototypes through participant ratings using the System Usability Scale 0-100 (0: unusable; 100: highly usable) [2], followed by a semi-structured interview where we asked about the relevance of the 32 statements and the preferences for the prototypes.

The statements were generally well-received, with most participants (12 out of 13) being able to comprehensively describe their AI systems' use. However, one participant found it more challenging, likely due to less experience in AI development. Despite that, the statements were considered accurate and applicable to a variety of AI uses. Participants rated the usability of the card stack prototype higher than the checklist (checklist: $\mu = 44.11$, $\sigma = 21.16$, while cards: $\mu = 66.43$, $\sigma = 16.01$). Therefore, for our final integration, the card-based interface was chosen.

By then analyzing the interviews' transcripts, we identified four design requirements used to revise our initial card-based interface (Figure 3). These requirements address *Role and Phase-based Selection* (R1), *Adequate Reading Time* (R2), *Skip Functionality* (R3), and *Input Sharing for Future Reviews* (R4).

To meet R1, we categorized statements by stakeholder roles (designers, engineers or researchers, managers or executives) [6] [11] and AI system phases (development, deployment, use) [12], and made sure the interface adjusts the number of relevant statements that are presented based on the selected role and phase. For R2, we presented statements as two-sided cards to ensure adequate reading time. For R3, we included a skip button for irrelevant statements and a progress tracker. Finally, to meet R4, we developed a summary page that compiles stakeholders' answers into downloadable PDF and JSON formats for easy sharing and review.

B. FillGen

This LLM-powered tool processes the previous stakeholders' input to pre-fill the impact assessment report. It does so by running four prompts in OpenAI's GPT-4[10]. We chose version 4 for its capability of interpreting legal documents and, more generally, for its top performance as of March 2024[11]. The four prompts are:

(1) SysInfoGen. The first prompt summarizes the input of the stakeholders (statements 1–27) to generate the description of the system's use in a five-component format suitable for risk assessment as per the EU AI Act [8]. The prompt has three parts: role, instructions, and output format. To begin with, the model is asked to take the role of a "Senior AI Technology Specialist" guiding it to generate a specific type of content. The instructions indicate, upon the stakeholders' input, to articulate the system's use into these five components. The first component is Domain. This specifies the industry or sector, like law enforcement. Purpose explains the goal, such as identifying potential crime hotspots. Capability describes the technology behind it, like object and action recognition. AI user is the one using the system, like the police force or security agencies.

[10] https://cdn.openai.com/papers/gpt-4.pdf
[11] https://arena.lmsys.org

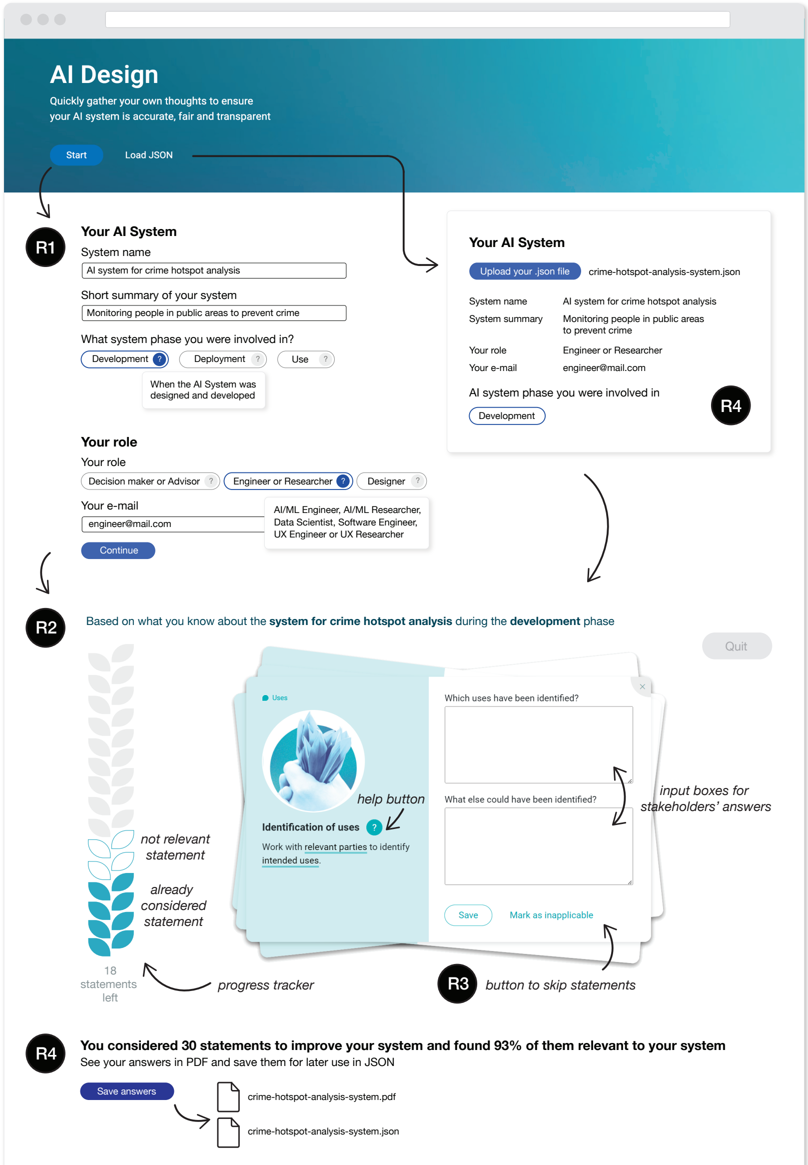
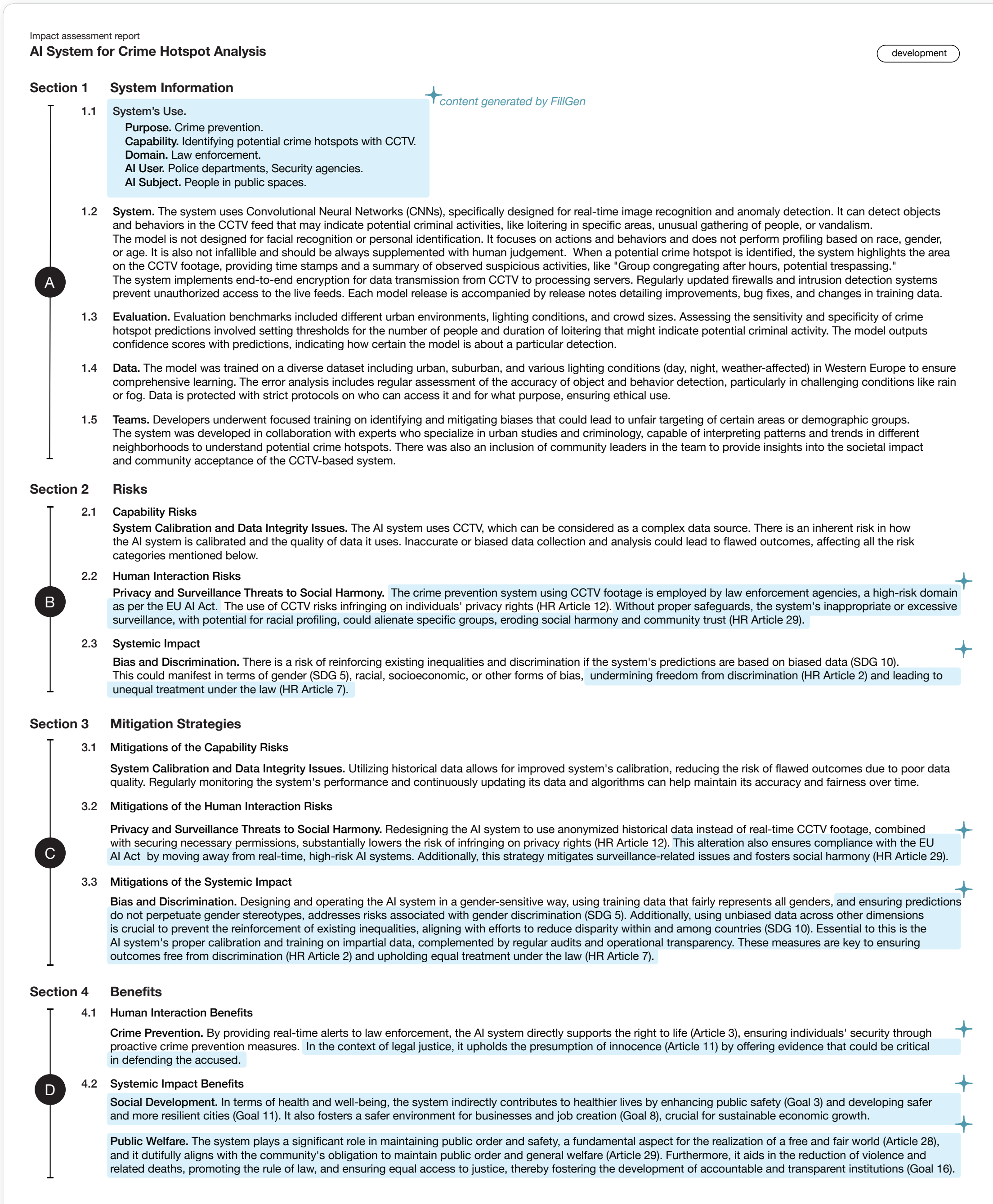


Figure 3: *StakeLinker's* final user interface is a one-page card stack. The interface meets four design requirements: (R1) adjusting the number of relevant statements based on the selected role and phase; (R2) presenting each statement as a card to provide adequate reading time; (R3) allowing to skip irrelevant statements and track the progress; and (R4) enabling sharing input for future reviews.

AI subject is the one affected by the system, such as people in public spaces. Then, the output format specifies that the description of the system's use in output (e.g., "crime hotspot analysis using CCTV") should be returned in the format [Domain, Purpose, Capability, AI user, AI subject] as, say, ["Law enforcement", "Identifying potential crime hotspots", "Object and action recognition", "Police departments, Security agencies", "People in public spaces"]. Finally, the output is put in the System Information section of the impact assessment report (Figure 4A), and propagated to the next prompt.

(2) RiskGen. The second prompt summarizes what the different stakeholders think about the risks of the system (statement 28) and finds more risks using guidelines from the EU AI



Impact assessment report

AI System for Crime Hotspot Analysis

development

Section 1 System Information

content generated by FillGen

1.1 System's Use.
Purpose. Crime prevention.
Capability. Identifying potential crime hotspots with CCTV.
Domain. Law enforcement.
AI User. Police departments, Security agencies.
AI Subject. People in public spaces.

1.2 **System.** The system uses Convolutional Neural Networks (CNNs), specifically designed for real-time image recognition and anomaly detection. It can detect objects and behaviors in the CCTV feed that may indicate potential criminal activities, like loitering in specific areas, unusual gathering of people, or vandalism.
The model is not designed for facial recognition or personal identification. It focuses on actions and behaviors and does not perform profiling based on race, gender, or age. It is also not infallible and should be always supplemented with human judgement. When a potential crime hotspot is identified, the system highlights the area on the CCTV footage, providing time stamps and a summary of observed suspicious activities, like "Group congregating after hours, potential trespassing."
The system implements end-to-end encryption for data transmission from CCTV to processing servers. Regularly updated firewalls and intrusion detection systems prevent unauthorized access to the live feeds. Each model release is accompanied by release notes detailing improvements, bug fixes, and changes in training data.

A

1.3 **Evaluation.** Evaluation benchmarks included different urban environments, lighting conditions, and crowd sizes. Assessing the sensitivity and specificity of crime hotspot predictions involved setting thresholds for the number of people and duration of loitering that might indicate potential criminal activity. The model outputs confidence scores with predictions, indicating how certain the model is about a particular detection.

1.4 **Data.** The model was trained on a diverse dataset including urban, suburban, and various lighting conditions (day, night, weather-affected) in Western Europe to ensure comprehensive learning. The error analysis includes regular assessment of the accuracy of object and behavior detection, particularly in challenging conditions like rain or fog. Data is protected with strict protocols on who can access it and for what purpose, ensuring ethical use.

1.5 **Teams.** Developers underwent focused training on identifying and mitigating biases that could lead to unfair targeting of certain areas or demographic groups.
The system was developed in collaboration with experts who specialize in urban studies and criminology, capable of interpreting patterns and trends in different neighborhoods to understand potential crime hotspots. There was also an inclusion of community leaders in the team to provide insights into the societal impact and community acceptance of the CCTV-based system.

Section 2 Risks

2.1 Capability Risks
System Calibration and Data Integrity Issues. The AI system uses CCTV, which can be considered as a complex data source. There is an inherent risk in how the AI system is calibrated and the quality of data it uses. Inaccurate or biased data collection and analysis could lead to flawed outcomes, affecting all the risk categories mentioned below.

B

2.2 Human Interaction Risks
Privacy and Surveillance Threats to Social Harmony. The crime prevention system using CCTV footage is employed by law enforcement agencies, a high-risk domain as per the EU AI Act. The use of CCTV risks infringing on individuals' privacy rights (HR Article 12). Without proper safeguards, the system's inappropriate or excessive surveillance, with potential for racial profiling, could alienate specific groups, eroding social harmony and community trust (HR Article 29).

2.3 Systemic Impact
Bias and Discrimination. There is a risk of reinforcing existing inequalities and discrimination if the system's predictions are based on biased data (SDG 10). This could manifest in terms of gender (SDG 5), racial, socioeconomic, or other forms of bias, undermining freedom from discrimination (HR Article 2) and leading to unequal treatment under the law (HR Article 7).

Section 3 Mitigation Strategies

3.1 Mitigations of the Capability Risks
System Calibration and Data Integrity Issues. Utilizing historical data allows for improved system's calibration, reducing the risk of flawed outcomes due to poor data quality. Regularly monitoring the system's performance and continuously updating its data and algorithms can help maintain its accuracy and fairness over time.

C

3.2 Mitigations of the Human Interaction Risks
Privacy and Surveillance Threats to Social Harmony. Redesigning the AI system to use anonymized historical data instead of real-time CCTV footage, combined with securing necessary permissions, substantially lowers the risk of infringing on privacy rights (HR Article 12). This alteration also ensures compliance with the EU AI Act by moving away from real-time, high-risk AI systems. Additionally, this strategy mitigates surveillance-related issues and fosters social harmony (HR Article 29).

3.3 Mitigations of the Systemic Impact
Bias and Discrimination. Designing and operating the AI system in a gender-sensitive way, using training data that fairly represents all genders, and ensuring predictions do not perpetuate gender stereotypes, addresses risks associated with gender discrimination (SDG 5). Additionally, using unbiased data across other dimensions is crucial to prevent the reinforcement of existing inequalities, aligning with efforts to reduce disparity within and among countries (SDG 10). Essential to this is the AI system's proper calibration and training on impartial data, complemented by regular audits and operational transparency. These measures are key to ensuring outcomes free from discrimination (HR Article 2) and upholding equal treatment under the law (HR Article 7).

Section 4 Benefits

4.1 Human Interaction Benefits
Crime Prevention. By providing real-time alerts to law enforcement, the AI system directly supports the right to life (Article 3), ensuring individuals' security through proactive crime prevention measures. In the context of legal justice, it upholds the presumption of innocence (Article 11) by offering evidence that could be critical in defending the accused.

D

4.2 Systemic Impact Benefits
Social Development. In terms of health and well-being, the system indirectly contributes to healthier lives by enhancing public safety (Goal 3) and developing safer and more resilient cities (Goal 11). It also fosters a safer environment for businesses and job creation (Goal 8), crucial for sustainable economic growth.

Public Welfare. The system plays a significant role in maintaining public order and safety, a fundamental aspect for the realization of a free and fair world (Article 28), and it dutifully aligns with the community's obligation to maintain public order and general welfare (Article 29). Furthermore, it aids in the reduction of violence and related deaths, promoting the rule of law, and ensuring equal access to justice, thereby fostering the development of accountable and transparent institutions (Goal 16).

Figure 4: Impact assessment report for an AI system identifying crime hotspots from CCTV footage. The report consists of four sections describing the system's: (A) use in the five-component format and information about system components, evaluation, data, and development teams; (B) risks; (C) mitigation strategies; and (D) benefits. Risks, mitigation strategies, and benefits are categorized in three layers: risks related to (model) capability, human interactions, and systemic impact [19]. Text with a light blue background and a star icon indicates new content generated by *FillGen* (30% of the report's content), and text without background indicates content provided by the stakeholders and summarized by *FillGen*.

Act, Human Rights (HRs),[12] and Sustainable Development Goals (SDGs)[13]. We chose these three documents because they match well with a DeepMind's framework that looks at three main areas [19]: the system's technical parts (called 'system's capability risks' in the framework), how people interact with it (human interaction risks), and its impact on society, the economy, and the environment (systemic risks). The EU AI Act focuses on the technical parts, Human Rights focus on how people interact with the system, and the Sustainable Development Goals look at its long-term impact on the world. *RiskGen*'s prompt has four parts: role, input, instructions, and output format. The role is a "Senior AI Technology Expert, specializing in compliance with the EU AI Act, SDGs, and HRs". The input includes parts of the EU AI Act about high-risk AI, the definitions of the 17 SDGs from the Sustainable Development Agenda, and the 30 articles from the UN Universal Declaration of Human Rights. The instructions involve using a step-by-step reasoning method (Chain-of-Thought) by breaking down the task into five smaller steps: 1) summarizing the risks identified by others; 2) writing a short description of how the system is used; 3) classifying the risk level of the system's use as either unacceptable, high-risk, or safe; 4) finding any extra risks that affect the SDGs or HRs because of how the system is used; and 5) organizing all the risks based on their importance to the system's technical parts, how people interact with it, and its impact on society. Finally, the output is then used to help fill out the Risk section of the impact assessment report (Figure 4B), and is also used in the fourth prompt called *MitigationGen*.

(3) BenefitGen. This third prompt gathers what the stakeholders think about the benefits of using the system (statements 31–32) and finds even more benefits. It works with four parts: role, input, instructions, and output format. The role is a "Senior AI Technology Expert, specializing in SDGs and HRs". The input includes the definitions of the 17 SDGs and the 30 articles from the UN Universal Declaration of Human Rights. The instructions again specify a Chain-of-Thought. Finally, the output is then used to help fill out the Benefits section of the impact assessment report (Figure 4D).

(4) MitigationGen. The fourth and last prompt gathers mitigation strategies from the stakeholders about how to reduce the risks of the system (statements 29–30) and suggests extra ways to lessen these risks, which were previously identified by *RiskGen*. It has three parts: role, instructions, and output format. The role is a "Senior AI Technology Expert, specializing in compliance with the EU AI Act, SDGs, and HRs". The instructions guide the model to arrange the strategies in a proper output format. This output is then used to pre-fill the Mitigations section of the impact assessment report (Figure 4C).

Before we used *FillGen* to make a full impact assessment report, we tested its four different prompts. The *SysInfoGen* prompt worked well because it turned input text into a standard format. However, *RiskGen* and *BenefitGen* sometimes repeated things they should not because they often referred to SDGs (Sustainable Development Goals) and Human Rights, and some of these overlap. For example, SDG 10, which is about reducing inequalities, and HR Article 2, which is about freedom from discrimination, ended up giving similar results. *MitigationGen* sometimes made up solutions or gave unclear ones. To solve these problems, we made some improvements to the prompts by: 1) having the model combine similar risks and benefits; 2) telling it not to suggest solutions if it was not sure; and 3) giving it examples of good-quality outputs for risks and benefits to learn from (few-shot learning).

Evaluating FillGen based on an impact assessment report.

To determine whether our framework works, *FillGen* was used to pre-fill a report for a hypothetical AI system that identifies crime hotspots from CCTV footage. *FillGen* summarized input from a developer and designer at our company who have experience with similar systems, and it also listed new risks, ways to fix them, and benefits. This AI system was deemed high-risk because it may indicate potential criminal activities in public places (Figure 4A), and it could invade

[12] https://www.un.org/en/about-us/universal-declaration-of-human-rights
[13] https://sdgs.un.org/2030agenda

privacy or increase racial inequalities (Figure 4B). However, these risks can be reduced if the system is trained with anonymous and diverse historical data, or if it is regularly checked (Figure 4C). By doing this, the system could help keep public order in a fair and democratic way (Figure 4D).

Evaluation. To test our pre-filled impact assessment report, we run a user study with 8 AI experts (3 women and 5 men, average age 33) who all had at least two years of experience in AI, including in areas like machine learning and natural language processing.

During the study, each person spent 10-15 minutes reading the report. After reading, they rated how they felt about the report's setup (how complete it was, how easy it was to use, and how applicable it was to other systems) and its details (how realistic the risks, solutions, and benefits seemed) using a scale from 1 (strongly disagree) to 5 (strongly agree). Then, we interviewed them to get more thoughts on the report's structure and content.

The results showed that most people thought the report was detailed ($\mu = 4.25$), applicable to many AI systems ($\mu = 4.5$), easy to use ($\mu = 4.6$), and realistic in presenting risks ($\mu = 4.5$), mitigation strategies ($\mu = 4.13$), and benefits ($\mu = 4.5$). They liked how the report was consistent and clearly described the system's use, especially when checking it against EU AI rules. As one participant stated, *"you see it [five-format component], and it sticks with you."*.

However, some thought the report was too general. For example, one participant said the solutions it suggested could really apply to any system handling personal data, not just camera-based AI. Another worry was that the report's polished look might make people accept it without enough critical review, missing the need for a deeper risk assessment. Overall, they agreed the report is a good starting point but still needs more detailed checks and input from legal experts to ensure accuracy and relevance.

We also have two more reports on our project's website[6]: one for a chatbot that checks responses in training tasks, and another for an AI system that evaluates damage after natural disasters. These are to show how the framework can be used in different scenarios.

DISCUSSION AND CONCLUSION

AI Design is a tool that helps developers and decision-makers create impact assessment reports about AI systems. It fills in parts of a report automatically, making it easier for developers to figure out if their AI is good to go and what problems they might need to fix. This is really useful for small companies because it saves time and money, helping them avoid big fines or costly fixes. The reports made can also be shared, helping everyone learn more about AI's impact.

AI Design is different from existing guidebooks for populating impact assessment reports in three main ways:

1) *Equal Importance:* Unlike some guides that focus too much on risks [15], our tool gives equal weight to the benefits and the risks.
2) *Personalized for Users:* Our tool adjusts its advice depending on who is using it, which means it is easier for everyone to understand and use, no matter their job role.
3) *Simpler and Practical:* Some guides are too long and hard to use, but our tool makes it straightforward with just 32 key statements to cover in a report.

From a technical point of view, our tool uses advanced AI to summarize important information and link it to complex rules like the EU AI Act. It also helps people understand how AI impacts society by including different viewpoints in the assessment process [13].

Practically, AI Design cuts down the costs and effort needed to meet legal standards. For example, when new data protection laws came in, some small companies struggled because they did not have enough resources[14]. Our tool helps prevent that by doing some of the heavy lifting, and it teaches people about the rules they need to follow.

However, there are a few limitations:

1) *Biases in Reports:* The reports might lean a certain way based on what stakeholders report. We need to find ways to check these biases and balance out the views. Methods like the Delphi method could help make sure no single viewpoint dominates the final assessment [17].

[14] https://gdpr.eu/2019-small-business-survey

2) *Not Covering All Impacts:* Our tool might miss some new or unique problems that need expert input. While it is a great starting point, experts need to check over the final report.
3) *Need for Good AI Governance:* Organizations should have steps in place to manage AI risks, involving everyone from designers to external auditors. One way to manage risks is to follow a four-step process [14]:

   First Step: Designers and developers make sure the AI meets ethical and technical standards. They give initial feedback on the AI's design.

   Second Step: An internal committee, like an Ethics Committee, checks the reports to make sure everything is right before they move on to the next stage.

   Third Step: The top management and other important groups within the company look at the report. They are the main people who need to understand and approve it.

   Fourth Step: External auditors review the report. This final check helps to make sure the AI follows all rules and to identify any problems that might still be there.
4) *Keeping Up with Changes:* Since our tool is based on the current EU AI Act, we need to keep it updated as new laws come out.

Overall, AI Design makes creating impact reports easier, but it still needs expert review and updates to stay effective and accurate.

REFERENCES

1. A. M. Barrett, D. Hendrycks, J. Newman, and B. Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. *arXiv preprint arXiv:2206.08966*, 2022.
2. J. Brooke. SUS: A "Quick and Dirty" Usability Scale. *Usability Evaluation In Industry*, 189(3):189–194, 1996.
3. Z. Buçinca, C. M. Pham, M. Jakesch, M. T. Ribeiro, A. Olteanu, and S. Amershi. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. *arXiv preprint arXiv:2306.03280*, 2023.
4. I. Cheong, K. Xia, K. J. K. Feng, Q. Z. Chen, and A. X. Zhang. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, page 2454–2469. ACM, 2024.
5. J. H. Choi. How to Use Large Language Models for Empirical Legal Research. *Journal of Institutional and Theoretical Economics (Forthcoming)*, 2023.
6. M. Constantinides, E. Bogucka, D. Quercia, S. Kallio, and M. Tahaei. RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles. In *Proceedings of the ACM on Human-Computer Interaction*, number CSCW, pages 1–28, 2024.
7. J. De Miguel Velazquez, S. Šćepanović, A. Gvirtz, and D. Quercia. Decoding Real-World AI Incidents. *IEEE Computer*, 2024.
8. D. Golpayegani, H. J. Pandit, and D. Lewis. To Be High-Risk, or Not To Be–Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, page 905–915, 2023.
9. M. Havrda and B. Rakova. Enhanced Well-Being Assessment as Basis for the Practical Implementation of Ethical and Rights-Based Normative Principles for AI. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 2754–2761. IEEE, 2020.
10. V. Herdel, S. Šćepanović, E. Bogucka, and D. Quercia. ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2024.
11. M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach. Co-designing Checklists to Understand Organizational Challenges and Opportunities Around Fairness in AI. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–14, 2020.
12. M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model Cards for Model Reporting. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, page 220–229, 2019.
13. H. Purohit, V. L. Shalin, and A. P. Sheth. Knowledge Graphs to Empower Humanity-Inspired AI Systems. *IEEE Internet Computing*, 24(04):48–54, jul 2020.
14. J. Schuett. Three Lines of Defense Against Risks From AI. *AI & Society*, 2023.
15. E. Sherman and I. Eisenberg. AI Risk Profiles: A Standards Proposal for Pre-deployment AI Risk Disclosures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38:21, page 23047–23052, 2024.
16. B. C. Stahl, J. Antoniou, N. Bhalla, L. Brooks, P. Jansen, B. Lindqvist, A. Kirichenko, S. Marchal, R. Rodrigues, N. Santiago, Z. Warso, and D. Wright. A Systematic Review of Artificial Intelligence Impact Assessments. *Artificial Intelligence Review*, 56(11):12799–12831, 2023.
17. B. Vasey, M. Nagendran, B. Campbell, D. A. Clifton, G. S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M. Ibrahim, X. Liu, B. A. Mateen, P. Mathur, M. D. McCradden, L. Morgan, J. Ordish, C. Rogers, S. Saria, D. S. W. Ting, P. Watkinson, W. Weber, P. Wheatstone, P. McCulloch, et al. Reporting Guideline for the Early-stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI. *Nature Medicine*, 28(5):924–933, 2022.
18. Z. J. Wang, C. Kulkarni, L. Wilcox, M. Terry, and M. Madaio. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–40, 2024.
19. L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, I. Gabriel, V. Rieser, and W. Isaac. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv preprint arXiv:2310.11986*, 2023.
20. B. Xia, Q. Lu, H. Perera, L. Zhu, Z. Xing, Y. Liu, and J. Whittle. Towards Concrete and Connected AI Risk Assessment (C^2AIRA): A Systematic Mapping Study. *arXiv preprint arXiv:2301.11616*, 2023.

Edyta Bogucka is a Research Scientist at Nokia Bell Labs Cambridge (UK). She works in the areas of data visualization, user experience design, and responsible AI. Contact her at edyta.bogucka@nokia-bell-labs.com

Marios Constantinides is a Senior Research Scientist at Nokia Bell Labs Cambridge (UK). He works in the areas of human-computer interaction, UbiComp, and responsible AI. Contact him at marios.constantinides@nokia-bell-labs.com.

Sanja Šćepanović is a Senior Research Scientist at Nokia Bell Labs Cambridge (UK). She works in the areas of social computing, earth observation, and responsible AI. Contact her at sanja.scepanovic@nokia-bell-labs.com.

Daniele Quercia is the Department Head at Nokia Bell Labs in Cambridge (UK) and Professor of Urban Informatics at King's College London. He works in the areas of computational social science, urban informatics, and responsible AI. Contact him at quercia@cantab.net.