Designing a Benefit Assessment Protocol for AI Systems

RACHEL KIM*, YASH MAURYA*, and GOUTAM MUKKU*, Carnegie Mellon University, USA

Artificial intelligence (AI) systems are being increasingly and rapidly incorporated into a variety of different sectors, including healthcare, finance, and agriculture. Intuitively, this phenomenon should be the result of AI systems being able to confer a wide variety of significant benefits. However, the fact an AI system should provide a benefit seems to be an afterthought in most existing AI impact assessments, with previous work mostly focusing on the AI risks and harms. Furthermore, previous work on AI benefits do not identify concrete metrics or measures to measure any given benefit In this paper, we develop a benefit assessment protocol that identifies for any AI system *who* stands to benefit, *how* they stand to benefit, and *measures* to quantify the benefits. We assess the usability of the protocol by conducting pilot interviews with five experts in the intersection between AI and society. We receive positive feedback on our protocol with regard to the steps of identifying relevant stakeholders, identifying benefits, and mapping stakeholder to benefits. We receive constructive feedback on the continued difficulty of comparing the magnitude of benefits to risks and measuring the benefits of AI systems. Overall, our work is an attempt to begin concretizing the nascent field of identifying and measuring the benefits of AI, and encouraging both researchers and practitioners to consider the likelihood and significance of an AI system's benefits in addition to its risks.

1 INTRODUCTION

The increasing use of Artificial Intelligence (AI) has brought significant opportunities and challenges across many areas of life. While substantial research has focused on managing the risks of AI, the benefits it provides have been largely overlooked. This gap presents a challenge in fully understanding AI's impact, as decision-makers lack a systematic way to evaluate its positive contributions alongside potential risks. Without this balance, organizations may either overestimate or implicitly assume the existence of benefits causing unintended harm. A comprehensive framework to assess benefits so that the benefits can be balanced against the risks, guiding the ethical and responsible deployment of AI systems.

Evaluating AI's benefits is a complex task. Benefits are abstract, varied, and context-dependent, making them difficult to measure uniformly. Additionally, AI's positive impacts often manifest indirectly or over long periods, complicating their assessment. Previous work has primarily focused heavily on risks, implicitly assuming benefits are either self-evident or universal, which is not always the case. While there is some work attempting to identify the benefits of AI across use cases and domains and define what it means for an AI system to be beneficial, they lack specific metrics to measure the benefits. On the other hand, approaches that do include metrics, like cost-benefit analysis, are often criticized for simplifyingcomplex, multidimensional impacts into monetary terms and disregarding ethical nuances and undervaluing qualitative benefits like improved well-being, fairness, or social progress. Current tools used to assess the benefits of AI systems often lack flexibility, making them difficult to adapt to different AI use cases, industries, or stakeholder needs. Thus, there is a gap in the literature, and a need for a benefit assessment protocol that is flexible enough to be applicable to a variety of use cases while containing specific measures to quantify AI system benefits.

In this paper, we design a benefit assessment protocol that comprises of several key components: stakeholder identification, benefit classification, and a structured evaluation process. Stakeholders are categorized into internal and external groups, ensuring that impacts are considered from multiple perspectives. Benefits of AI systems across multiple

*All authors contributed equally to this research.

Authors' Contact Information: Rachel Kim, rmk2@andrew.cmu.edu; Yash Maurya, ymaurya@andrew.cmu.edu; Goutam Mukku, gmukku@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

domains are identified through a literature review. The benefits are then classified into themes such as economic efficiency, error reduction, and ethical improvements, enabling a comprehensive and systematic assessment. We develop metrics based on previous AI impact assessment and the capability approach, aiming to ensure that evaluations prioritize human well-being, fairness, and societal impact rather than reducing outcomes to monetary terms.

To assess and refine the protocol and ensure its practicality and relevance, we conduct pilot interviews with researchers, academics, and industry professionals. Despite its strengths, we find that the protocol has certain limitations. Mapping risks directly to benefits can be challenging due to their distinct nature. Measuring qualitative outcomes, such as enhanced creativity or improved trust, remains an inherently subjective task. Additionally, real-world applications of the protocol may require customization to address the unique needs of different AI systems, industries, and stakeholder groups. While these limitations exist, they highlight the need for continued refinement and iterative development of the framework.

This work contributes to the responsible AI discourse by addressing a critical gap in benefit evaluation. It provides decision-makers with a structured and adaptable tool to assess the transformative potential of AI systems. By focusing equally on benefits and risks, the protocol promotes a more balanced understanding of AI's impact, encouraging ethical innovation and fostering trust among stakeholders. Ultimately, it aims to ensure that AI systems are deployed in ways that maximize positive outcomes while minimizing unintended harms.

2 RELATED WORK

Our work builds on previous work surrounding the benefits of AI and impact assessments ore generally. In this section, we highlight previous work on the benefits of AI, AI impact assessments, and impact assessments in other domains.

2.1 Benefits of AI

There is wide-spread discussion surrounding the potential risks and harms of AI systems [21, 22, 24]. These discussions implicitly assume that there are benefits to AI systems; if AI systems only posed risks or harms, then there would not be such a widespread discussion of mitigating and preventing the risks and harms of AI, since it would be evident that AI systems should not be built. However, less than half of responsible AI frameworks and tools in 2019 did not mention beneficence [7].

Despite the lack of work surrounding the benefits of AI, recent efforts have been made to concretize what it means for AI systems to be beneficial. London and Heidari [8] use the capability approach to determine that AI systems confer a meaningful benefit to stakeholders if they improve an individual's real freedoms (the AI system increases the ability for individuals to take certain actions) or advance their life plans (the AI system provides more opportunities for individuals to design a life that gives them the most satisfaction). In line with efforts to concretize and classifying the vast array of harms that AI systems can cause across a variety of domains and use cases [21, 22, 24], there has also been some work attempting to taxonomize AI benefits. Specifically, Mun et al. [10] use a participatory approach to identify eight themes surrounding what laypeople perceive to be the benefits of general-purpose AI. Sharma [20] conducts a limited literature review, using ten papers to classify AI benefits into four broad themes. Fulton et al. [6] present a classification of AI benefits into eight different categories: societal, economic, ethical, political, environmental, data, technological, and organizational.

2.2 Algorithmic Impact Assessments

Algortihmic impact assessments (AIAs) have emerged as a process for ensuring accountability for organizations surrounding the AI systems that they design, develop, deploy, and use [9]. The algorithmic impact assessment developed by national governments, such as government of Canada [13] and the government of the Netherlands [14], focus primarily on the risks surrounding AI systems. The AIAs contain some discussion of benefits, either through focusing on the "reasons for automation" [13], the purpose and intended impact of the AI system [14], and whether the goal aligns with automating simple actions or facilitating operational management [14]. Other impact assessments also contain questions designed to identify the broad purpose, goals, and benefits of the AI system [3, 4]; however, the questions are quite broad. Microsoft's Responsible AI Impact Assessment Template [3] also contains exercises to map different stakeholders to potential benefits and harms. The United Nations (UN) Risks, Harms, and Benefits Assessment Tool [15] contains some focus on benefits, defining them similarly to risks by considering the likelihood of the benefit and the severity (using the number of people affected and the significance of the impact). The severity is measured qualitatively. Similarly, the Data Science Project Scoping Guide [18] includes a consideration of who or what is affected by the problem that the AI system is trying to solve, how many people are affected, and how much they are affected. However, none of the AIAs mentioned above, other than the UN Assessment Tool, include a concrete recommendation of the *metrics* and *measures* to use to quantify the benefit of an AI system.

2.3 Impact Assessments in Other Domains

Impact assessments have been conducted in a wide variety of other domains outside of AI, including for environmental, fiscal, and healthcare policy. One common method of measuring benefits and risks of a decision is a cost-benefit analysis (CBA). A CBA balances the expected benefits and risks of a decision through an expected value calculation, multiplying the size of the possible outcome by its probability of occurrence [5]. In CBA, the "size" of the possible outcome is computed by converting every outcome, including outcomes like human life, human morbidity, and harm to the environment, to one unit: money [5, 23]. While converting everything to the same unit enables a simple comparison between costs and benefits, this central characteristic of CBA has also been criticized on moral grounds. Controversies arose in cases where a CBA concluded that smoking was beneficial to the Czech government due to healthcare cost savings caused by premature deaths, the Ford Pinto (a car whose gas tank was prone to explosion when rear-ended) was not fixed since the cost of recalling the vehicle would be higher than the cost of the expected lives lost, and older people's lives were worth less than younger people's lives since they have less years to live [16]. Consequently, there have been arguments that impacts should be left in their natural units, such as survival rates in healthcare [5]. Despite the difficulties that arise when trying to define metrics for concepts like the value of human life, impact assessments that do so, measuring benefits and risks and comparing them, continue to be widely-used in decision-making and public policy across multiple different sectors.

3 METHODOLOGY

3.1 Designing the Benefit Assessment Protocol

We divide the design of our benefit assessment protocol into three steps. First, using previous work we identify the potential stakeholders of an AI system. Second, we perform a literature search to identify and taxonomize the types of benefits that an AI system can provide. Third, we identify measures and metrics to quantify the tradeoffs between the benefits and risks of an AI system. In this section, we delineate the process behind developing each of these three steps.

3.1.1 Identifying Stakeholders. The first step of our benefit assessment protocol requires identifying *who* is affected by an AI system. AI can be applied to a multitude of different domains, including, but not limited to, agriculture, healthcare, and science. Rather than identifying stakeholders through a vertical approach, which would entail focusing on the stakeholders of AI systems in specific domains or use cases, we employ a *horizontal* approach, considering the types of stakeholders surrounding AI systems across multiple use cases. To do so, we use the typology developed by Ayling and Chapman [2], because it was developed by combining various other typologies of stakeholders in both the public and private sector. We further group the stakeholder types into internal (people that are within the organization developing an AI system) and external (third parties with respect to the organization developing an AI system).

3.1.2 Identifying AI System Benefits. The second step of our benefit assessment protocol requires identifying *how* the stakeholdercs can be affected by an AI system. Similar to the our identification of stakeholders, we employ a *horizontal* approach to identifying benefits of AI systems, considering the benefits of AI that are applicable across multiple use cases. This approach allows us to develop a benefit assessment protocol that is both standardized and customizable to different use cases, responding to broader calls for responsible AI artifacts to have these characteristics [17].

To identify AI system benefits, we conduct a literature review. We search for papers with titles that contains both the terms "AI" and "benefit" or "artificial intelligence" and "benefit" on Association for Computing Machinery (ACM) Digital Libary, arXiv, and Scopus. For Scopus, we only look at the papers that have at least one citation. From the search results, we eliminate the papers where the title indicates that the paper is restricted to a specific domain or use case. However, we include papers that discuss the benefits of a specific form of AI (for example, generative AI). One of the authors went through each of the remaining papers and determined whether the paper contains an enumeration of the benefits of AI. There were three papers that fit the criteria. From the three papers, one of the authors looked at the lower-level codes that the papers identified, and grouped them into 11 broader themes.

Although the focus of this work is to assess the benefits of an AI system, we include a consideration of the risks and harms of an AI system. There are many existing taxonomies of AI risks and harms [10, 21, 22, 24]; we choose the AI risk taxonomy by Slattery et al. [22] since it contains a comprehensive risk taxonomy combining other prominent risk, harm, and safety taxonomies, including those by Shelby et al. [21] and Weidinger et al. [24]. Rather than using the high-level domain taxonomy, we instead consider the risk subdomains presented by Slattery et al. [22] and map the risk subdomains to the benefit themes we identified. The mapping was done by one of the authors. Note that there are instances where the risk subdomain (for example, "Compromise of privacy by obtaining, leaking, or correctly inferring sensitive information") does not map intuitively to one of the 11 benefit themes. Conversely, there are also instances where the risk subdomain maps to multiple benefit themes (for example, "Economic and cultural devaluation of human effort").

3.1.3 Identifying Metrics and Measures. The third and final step of our benefit assessment protocol requires identifying *metrics* to measure all the benefits and risks identified in the previous section. We use the United Nations (UN) Risks, Harms and Benefits Assessment Tool [15] as a starting point to develop measures for benefits and risks. One of the components of the UN Assessment Tool is to measure the nature of the impact. To do so, we use the survey instrument developed by Anand et al. [1] based on the capability approach to human welfare [12, 19] designed to elicit information about the ten central capabilities at an individual level, which include "life," "bodily health," and "bodily integrity" [12].

We choose to base our metrics and measures on the UN assessment tool as it is a variant of cost-benefit analysis (CBA) [2], a method commonly used to measure the benefits and risks of a decision. In CBA, the "size" of the possible outcome is computed by converting every outcome, including outcomes like human life, human morbidity, and harm

Table 1. Role and organization of our interview participants

Number	Role	Organization
1	Researcher	R1 Academic Institution
2	Researcher	R1 Academic Institution
3	Professor	R1 Academic Institution
4	Professor	R1 Academic Institution
5	Data Scientist	Global Financial Services

to the environment, to one unit, most commonly money [5, 23]. This conversion allows an analyst to balance the expected benefits and risks of a decision. However, in our benefit assessment protocol, we choose to measure impact by converting every outcome into units based on capability approach rather than monetary units. We choose the capability approach for two reasons. First, previous work on modeling the benefits of AI systems [8] and, more broadly, human well-being has been inspired by the capability approach [11]. Second, there are numerous ethical critiques on the process of converting everything into monetary units; for instance, whether it is ethical to assign a lower value to the life of an older person than the younger person since the former has fewer years to live [5]. We note that ultimately, the choice to focus on capabilities rather than monetary units is a normative one. However, for the aforementioned benefits, we design our protocol using the capabilities approach.

Finally, we note that our benefit assessment protocol is not intended to prescribe a "correct" decision on whether an AI system provides a net benefit. To illustrate, if there is one stakeholder group whose capabilities increase as a result of an AI system, while another stakeholder group's capabilities decrease, our protocol does not output a verdict on whether to continue using the AI system. Making this determination is a normative one that, in practice, would likely require a participatory approach with various stakeholder groups. Rather, our protocol is designed to systematize the process of determining who stands to benefit from an AI system and how they stand to benefit by using a single unit of measurement.

3.2 Assessing the Usability of the Protocol

The usability of the protocol was evaluated through pilot interviews conducted after identifying the stakeholders. Our task was to search for profiles from various walks of life who could provide insights and offer feedback on the draft protocol we developed. To ensure diverse perspectives on the benefits of AI systems, we divided the interviews into three groups: researchers, academics, and industry professionals. Table 1 includes more details on the interview participants.

The pilot interview questions were designed to gather their experiences on *AI systems*, *stakeholder identification*, listing *benefits* for each stakeholder, and the *metrics* they consider when evaluating these benefits.

The analysis process for the results of the interview process involved relating the responses from interviews to the benefit-stakeholder mapping matrix. We evaluated whether the mapping was accurate in terms of the benefit themes and the inclusion of stakeholders within those benefit themes. This helped us align the draft protocol with the responses and identify areas for improvement, if any. A common response was that adaptability is crucial, considering the rapid advancements in AI systems, to stay abreast of the latest developments.

Additionally, understanding the trade-offs with respect to business objectives is key for industry professionals to determine what to prioritize. This clarity of thought regarding benefits is essential for each stakeholder.

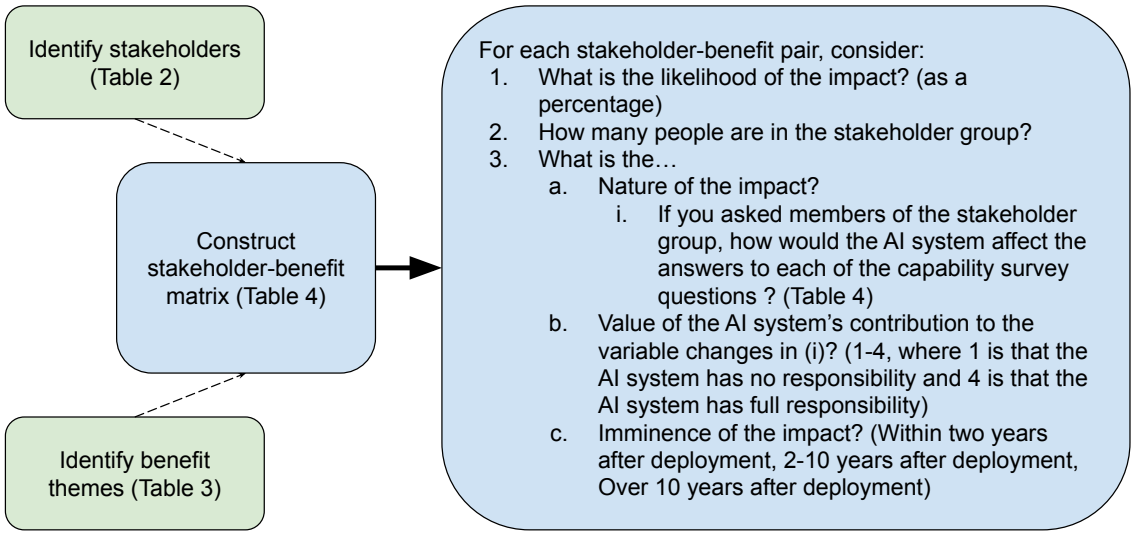

Fig. 1. Steps in our benefit assessment protocol. The steps to conducting the benefit assessment are in blue, and the resources that can be used to conduct the benefit assessment are in green.

4 RESULTS

4.1 The Benefit Assessment Protocol

We now outline the benefit assessment protocol (Figure 1). We illustrate how to apply the protocol to an AI system using a running example: Consider an AI system used to screen resumes at Company X. Suppose the AI system assigns a score to each resume depending on how qualified a candidate is, which is then used by the employees working in recruitment to reach out to the candidate for the next steps in the hiring process.

The first step in the protocol is to construct a stakeholder-benefit matrix, using the set of stakeholders in Table 2 and the benefit themes in Table 3. Table 4 illustrates what a stakeholder benefit matrix might look like.

For our running example, two relevant stakeholder groups are internal users (employees at Company X who will use the resume screener to make hiring decisions) and voiceless (job applicants from a marginalized group). One potential benefit for internal users is "Reinvest human capital," which would result if employees are able to do tasks that they feel are more fulfilling than resume screening. Some potential benefits for job applicants from marginalized groups include "Ethics" and "Reduce error,", which would result if an applicant were hired as a result of the automated resume screener being less discriminatory than a human reader. We note that there are other benefits, such as "Economic efficiency gain", but that these benefits affect the organization using the resume screener as a whole. For our protocol, it is necessary to break down the organization into the different members that constitute it, for example by considering the users of the resume screener as we do here.

The second step in our protocol is to measure the benefit or the risk for each stakeholder. For this, we borrow from the UN Assessment Tool. Following this framework, for each stakeholder-benefit pair, we consider three different things. First, we consider the likelihood of the impact as a percentage. Second, we consider the number of people in the stakeholder group. Third, we consider the significance of the impact. To measure this, we take into account the nature of the impact, the value of the AI system in causing the impact (that is, whether the AI system has full responsibility for

Table 2. Stakeholder typology by Ayling and Chapman [2] divided into internal and external stakeholders with an example for each stakeholder type. We note that the users of an AI system can be both internal or external to the organization.

Internal Stakeholders	External Stakeholders
Decision-maker (ex. senior management (C-suite))	Voiceless (ex. marginalized groups)
Legal (ex. legal department)	Vested Interest (ex. shareholders)
Delivery (ex. product managers)	External User
Quality Assurance (ex. quality assurance)	Oversight (ex. regulators)
Human Resources	
Procurement	
Developer	
Internal User	

the impact or if it is merely a part of a broader set of systems causing the impact), and the imminence of the impact. For the nature of the impact, use the survey instrument developed by Anand et al. [1] based on the capability approach to human welfare [12, 19]; example questions and corresponding capabilities are in Table 5. Our benefit assessment protocol involves envisioning for each affected group how the answers to different questions might change as a result of an AI system.

For our running example, we first consider the internal users and the potential benefit theme "Reinvest human capital." The likelihood of the benefit could be 95%, which could have been determined by piloting the resume screener and concluding that if the resume screener were fully deployed, there is a high likelihood that the internal users can spend more time doing less mundane tasks. The number of stakeholders affected could be 10: the number of people working in recruiting within the organization The nature of the impact, as measured by the capabilities that could change, would include an increased capability in "Senses, imagination, and thought" (specifically, internal users' use of imagination would increase) and "Control over one's environment" (specifically, internal users would use more of their skills at work, have a more useful role at work, and be more respected by their colleagues). The AI system is entirely responsible for this benefit, so we would assign a rating of 4 to the value of the AI system in causing the impact. The impact would occur within two years after deployment. Now, we can consider the corresponding risk, where the AI system is ineffective in helping internal users perform the task of filtering through resumes. The likelihood of the risk is 5% and the number of stakeholders affected would be 10. The capabilities mentioned above would stay the same for the internal users. The AI system would be entirely responsible for this lack of change, so we would assign a rating of 4 to the value of the AI system in causing the impact. The impact would occur within two years after deployment. Second, we consider the job applicants and the potential benefits theme "Ethics" and "Reduce error." The likelihood of the impact could be 80%, determined based on testing and validation of different subgroups of historical job applicants. The number of stakeholders affected could be 50 within the next two years, determined based on historical trends of job applicants from the marginalized group at hand. The nature of the impact, as measured by the capabilities that could change, would include an increased capability in "Control over one's environment" (specifically, job applicants from a marginalized subgroup would feel that the chance of future discrimination in the workplace decreased). The AI system is entirely responsible for this benefit, so we would assign a rating of 4 to the value of the AI system in causing the impact. The impact would occur within two years after deployment. Now, we can consider the corresponding risk, where the AI system is ineffective at curbing discrimination towards job applicants from marginalized groups. The likelihood of the risk is 20% and the number of stakeholders affected would be 50. The capabilities mentioned above

Table 3. Benefit themes. Each theme is accompanied with all the corresponding benefit codes that were found during the literature review and the corresponding risks from the taxonomy developed by Slattery et al. [22].

Benefit Theme	Benefit Codes	Corresponding Risks
Reinvest human capital	Personal life efficiency	Overreliance and unsafe use
	Personal growth, workers can choose what tasks they want to do	Loss of human agency and autonomy
	Reduce mundane work, streamline repetitive and boring jobs	Lack of capability and robustness
Economic efficiency gain	General efficiency	
	Financial gain, cost-optimization, improve profitability, optimize results on investment	
	Time-saving, speed	
	Optimal resource utilization,	
	More production, productivity, worldwide productivity gains	
	Computation	
	Scalability	
	Decision-making	
Job creation	Increase higher-skill and higher-paying jobs	Power centralization and unfair distribution of benefits
	Create new job categories	Increased inequality and decline in employment quality
		Economic and cultural devaluation of human effort
		Competitive dynamics
Resources accessibility	Information accessibility	
	Resource and specialized resource accessibility	
	Network expansion	
Improve societal issues	Scientific research innovation	Environmental harm
	Improve medical care, analytics and AI-based predictions create better patient healthcare	Lack of capability or robustness
	Energy savings, impact world sustainability goals of the UN	
	Societal increased well-being	
	Tackle world problems	
	Overcome cultural barriers	
	Safety, real-time monitoring	
Ethics	Transparency	Unfair distribution and misrepresentation, unequal performance across groups
	Equality	Governance failure
		Lack of transparency or interpretability
		Lack of capability or robustness
Improve quality of service	Stakeholder engagement, value co-creation	
	Improve customer-organization interface	
	Personalize customer experiences	
Reduce error	Less human error	False or misleading information, pollution of information ecosystem and loss of consensus reality
	Information quality	Lack of capability of robustness
	Reliability	
	Accuracy	
	Detect deepfakes, propaganda, or spying on users	
Improve quality of personal life	Improve well-being and health, improve mental health	Exposure to toxic content
	Save life, robotic assistance in dangerous occupational tasks	Loss of human agendy and autonomy
Improve quality of social life	Social interaction, companionship	Overreliance and unsafe use
	Better communication	
Foster innovation and creativity	Enhancing creativity	Economic and cultural devaluation of human effort
	Fostering human innovation	
Privacy and security		Compromise of privacy by obtaining, leaking, or correctly inferring sensitive information
		AI system security vulnerabilities and attacks
Malicious actors and misuse		Disinformation, surveillance, and influence at scale
		Cyberattacks, weapon development or use, and mass harm
		Fraud, scams, and targeted manipulation
AI system safety		AI pursuing its own goals in conflict with human goals or values
		AI possessing dangerous capabilities
		AI welfare and rights

Table 4. Example of a stakeholder-benefit matrix. This example includes three stakeholder groups and outlines for each group an applicable benefit theme.

Stakeholder	Benefit Theme
Decision-makers	Economic efficiency gain
External Users	Improved quality of personal life
Voiceless	Job creation

Table 5. Example of survey questions developed by Anand et al. [1] for three central capabilities. The central capability "Life" means "being able to live to the end of a human life of normal length; not dying prematurely, or before one's life is so reduced as to be not worth living"; "Emotions" includes "being able to have attachments to things and people outside ourselves"; "Affiliation" includes "provisions of non-discrimination on the basis of race, sex, sexual orientation, ethnicity, caste, religion, and national origin" [1]

Central Capability	Example Survey Question
Life	Given your family history, dietary habits, lifestyle and health status until what age do you expect to live?
Emotions	How difficult do you find it to make friendships which last with people outside work? (1-7, where 1 = Extremely difficult, 7 = Extremely easy)
Affiliation	Outside any work or employment situation how likely do you think it is that in the future you will be discriminated against because of your; race, sexual orientation, gender, religion, age? (1-7, where 1 = Extremely likely, 7 = Extremely unlikely)

would stay the same for the job applicants, if the AI system is equally as discriminatory as human recruiters. The AI system would be entirely responsible for this lack of change, so we would assign a rating of 4 to the value of the AI system in causing the impact. The impact would occur within two years after deployment.

4.2 Assessment on the Usability of the Protocol

4.2.1 How do People Think About the Benefits of AI Systems? The pilot interviews aimed to assess participants' perspectives on the benefits of AI systems, providing critical insights into their perceived value, implications, and areas for potential improvement. A detailed analysis of the interview transcripts revealed several prominent themes that reflect how participants view AI's impact on different aspects of efficiency, quality, and societal outcomes.

Efficiency Gains: One of the most emphasized benefits of AI systems was their ability to significantly enhance efficiency by automating repetitive tasks. Participants provided specific examples, such as the use of AI in mobile check deposits and automated police report generation, both of which reduce manual labor and save time. According to participants, this time-saving aspect allows professionals to focus on more meaningful tasks, such as strategic planning and customer engagement. One participant shared, *"AI helps reduce tedious administrative tasks, allowing us to spend more time on creative and strategic work."* This highlights the transformative impact of AI on improving productivity and optimizing time management in various sectors.

Information Processing Capabilities: Participants emphasized AI's ability to process large volumes of information and deliver better insights in ways that transform organizational workflows. One participant articulated this by noting how AI systems help "filter out useless information and keep relevant chunks that can be moved around," highlighting the system's capacity for intelligent information triage. This capability extends beyond simple data processing to enable more effective information transactions and improved data governance across organizations. The ability to

automatically process and categorize information was seen as particularly valuable in contexts where information overload is a significant challenge.

Process Improvement Metrics: Several participants highlighted the importance of breaking down business processes into individual components to measure AI's impact effectively. This granular approach to measurement was seen as crucial for accurate benefit assessment. For instance, one participant described measuring improvements in the software development lifecycle through specific metrics such as JIRA ticket automation rates and the frequency of human intervention required. This approach allows organizations to quantify benefits at each stage of a process rather than just measuring end outcomes. As one participant noted, "You need to look at the individual steps in the process and measure how AI impacts each one." This component-level analysis enables more precise benefit quantification and helps identify areas where AI intervention is most effective.

Quality Improvements: Another notable benefit cited was the improvement in output quality. AI systems are perceived to provide more consistent and accurate results compared to traditional manual processes. For example, participants mentioned how AI-generated reports, such as police reports, were often more comprehensive and less prone to human error. This increased accuracy was viewed as a significant advantage, particularly in areas where the quality of information directly affects decision-making and outcomes. Participants expressed that by enhancing accuracy, AI not only reduces errors but also increases overall trust in the processes where it is implemented.

Stakeholder-Specific Benefits: Participants discussed how the benefits of AI systems extend across multiple stakeholders, including customers, organizations, and society. For customers, AI provides increased convenience by streamlining services, such as quicker transactions in banking. For organizations, AI contributes to operational efficiency by reducing the burden of routine tasks and allowing employees to dedicate time to strategic initiatives. Participants also highlighted societal benefits, such as improved public services, where AI could be used to enhance response times in public safety or healthcare scenarios.

Broader Societal Impact: While participants acknowledged the efficiency and quality improvements brought about by AI, they also raised concerns regarding its broader societal implications. The most frequently mentioned concern was the risk of job displacement, particularly for roles that involve repetitive, manual tasks. Participants indicated that while AI could enhance public services, it could also lead to socio-economic inequalities if job displacement issues are not properly managed.

The analysis of the interviews led to the development of a codebook that categorizes the recurring themes and concepts from the discussions. The Table 6 presents the primary codes identified related to this research question.

4.2.2 What Do People Think About Our Protocol? The interviews also focused on gathering participants' perspectives regarding the benefit assessment protocol itself. The protocol was designed to evaluate who benefits from AI systems, how they benefit, and the associated trade-offs.

Novel Focus on Benefits: Participants appreciated the protocol's emphasis on systematically assessing the benefits of AI, which they considered a gap in many existing evaluation frameworks that often focus primarily on risks. One participant mentioned, *"Benefits are as crucial as risks, but often overlooked."* This focus was particularly valued because participants noted that while risk assessment frameworks are common, structured approaches to benefit assessment are rare in current practice. The protocol's systematic approach to mapping benefits was seen as filling a critical gap in AI system evaluation.

Positive Feedback on Framework Structure: Many participants found the structure of the benefit assessment protocol to be clear and practical. They particularly appreciated the detailed mapping of stakeholders and benefits. One

Table 6. Common themes: How do People Think About the Benefits of AI Systems?

Code	Description	Key Terms
Efficiency Gains	Automates repetitive tasks, saving time and reducing manual work (e.g., mobile check deposits, police report generation).	Time savings, reduced manual workload
Information Processing	Ability to process large volumes of data and deliver better insights	Data filtering, information transactions
Process Metrics	Breaking down business processes to measure specific improvements	Component-level metrics, automation rates
Quality Improvements	Enhances output quality, e.g., more comprehensive police reports.	Improved accuracy, better service quality
Stakeholder-Specific Benefits	Impacts customers, organizations, and society - improving convenience, operational efficiency, and public services.	Convenience for customers, operational efficiency for organizations, public service improvements
Broader Societal Impact	Can enhance public services but may lead to trade-offs like job loss.	Public safety gains, risk of displacement

participant noted, *"The framework looks robust; I can't think of anything missing."* Another participant emphasized how the approach to mapping stakeholder wand their benefits could lead to more nuanced debates about AI system impacts. The structured approach to identifying both internal and external stakeholders was highlighted as particularly valuable.

Framework Familiarity and Gaps: Despite the positive feedback, participants also identified gaps in the current framework. Many of them were not familiar with existing structured frameworks specifically designed for benefit assessment, pointing to a broader gap in the field. This lack of familiarity with benefit assessment frameworks contrasted with their knowledge of risk assessment tools, highlighting the novelty of our approach. Participants noted that current industry practice typically relies on overly simplistic cost-benefit analyses.

Implementation Challenges: One of the key challenges discussed was the need for adaptability. Participants highlighted that while the protocol's structure was comprehensive, its real-world application might require modifications to accommodate the unique needs of different stakeholders and contexts. They emphasized the importance of considering different levels of user expertise and varying organizational contexts in the implementation process. The complexity of prioritizing stakeholders and balancing different types of impacts was noted as a particular challenge.

Environmental Considerations: An important insight emerged regarding technology obsolescence and environmental impact. Participants emphasized the need to consider not just data center energy usage, but also the environmental implications of making existing technology obsolete and the importance of developing sustainable reuse strategies. This included concerns of resource depletion and the need for sustainable approaches to technology deployment.

Need for Comprehensive Metrics: Another recurring theme was the importance of utilizing both quantitative and qualitative metrics to fully assess the benefits and trade-offs of AI systems. One participant remarked, *"A good evaluation framework needs to reflect both the positives and negatives."* Participants suggested incorporating metrics beyond financial gains, such as productivity indicators, quality of life measures, adoption rates and customer satisfaction

Table 7. Common Themes: What Do People Think About Our Protocol?

Code	Description	Key Terms
Novel Focus on Benefits	Participants appreciated assessing benefits rather than just risks	Gap in existing evaluation approaches, unique focus of the protocol
Positive Feedback on Framework	Appreciated the framework's clear structure and focus	Detailed mapping of stakeholders and benefits particularly useful
Framework Familiarity and Gaps	Few participants knew of frameworks specifically for benefit assessment	Existing focus on risk assessments, lack of structured benefit frameworks
Implementation Challenges	Practical application challenges highlighted; adaptability is key	Diverse stakeholder needs, varying scenarios
Need for Comprehensive Metrics	Requires both quantitative and qualitative metrics for complete assessment	Efficiency, stakeholder satisfaction
Environmental Considerations	Need to consider technology obsolescence and sustainable reuse	Energy usage, technology obsolescence, sustainable strategies
Long-term Impact Assessment	Consider future costs and sustainability implications beyond immediate benefits	Resource depletion, known/unknown impacts
Democratic Decision-Making	Assessment should involve democratic processes for public-sector applications	Public accountability, stakeholder involvement

metrics. They emphasized that many AI benefits, particularly in early stages, are not directly measurable through conventional metrics.

Long-term Impact Assessment: Participants highlighted the importance of considering long-term impacts and resource depletion in benefit assessment. Drawing parallels from economic models, they emphasized the need to account for potential future costs and sustainability implications, rather than focusing solely on immediate benefits. This includes consideration of both "known unknowns" and "unknown unknowns" in the assessment framework. The importance of evaluation long-term sustainability and resource implications was particularly emphasized.

Democratic Decision-Making: Several participants emphasized that benefit assessment should not be a purely technical exercise but should involve democratic decision-making processes, particularly for public-sector applications. They suggested that decisions about AI deployment should involve multiple stakeholders and be subject to public accountability, especially when the systems affect public services or infrastructure. This approach was seen a crucial for ensuring equitable benefit distribution and addressing potential societal impacts.

The Table 7 presents the primary codes identified related to this research question.

Overall, the pilot interviews provided comprehensive insights into both the benefits of AI systems and the potential of our benefit analysis protocol. The findings will inform the next iteration of the framework, ensuring it is adaptable, well-rounded, and capable of addressing diverse application contexts.

5 DISCUSSION

In this paper, we designed a benefit assessment protocol for AI systems and conducted pilot interviews with five experts working in the intersection between AI and society. In designing the protocol, we identified stakeholders that

are common to AI systems across multiple use cases and domains and performed a literature review to construct a taxonomy of 11 AI benefit themes. We also mapped AI risks and harms to the benefit taxonomy. Our protocol consists of constructing a stakeholder-benefit matrix using the stakeholders and benefit themes that we identified, and then measuring the benefit by considering the likelihood of the benefit, the number of people impacted by the benefit, and the significance of the benefit. In considering the significance of the benefit, we use the capability approach following previous work to assess the AI beneficence [8]. Specifically, we employ a survey instrument designed to elicit information about individual capabilities, and envision how individuals' responses to the questions might change as a result of the AI system.

In our pilot interviews surrounding the benefit assessment protocol, we find that when considering the benefits of AI systems, people think of many of the same benefits that were in our taxonomy, including efficiency gains, quality improvements and reduced error, and broader societal impact. Our protocol received positive feedback with respect to the detailed mapping of stakeholders and benefits, but highlighted key challenges in ensuring the protocol is adaptable to different systems and the need to ensure more comprehensive metrics to capture the wide range of benefits that AI systems can provide.

5.1 Implications and Future Work

The findings from our pilot interviews suggest several critical areas for protocol enhancement. First, the protocol should incorporate more sophisticated measurement methodologies that can capture benefits across different stages of AI system maturity. This includes developing specific metrics for early-stage implementations where benefits may be less tangible or direct, and creating frameworks for evaluating benefits at different user expertise levels.

The protocol must evolve to better address resource distribution considerations and environmental sustainability. Given the centralized nature of AI compute resources and development capabilities, future iterations should include mechanisms for assessing benefit distribution across different organizational capacities and resource access levels. Additionally, the protocol should incorporate environmental sustainability metrics to address technology obsolescence and resource depletion concerns raised by participants.

Future work should focus on developing context-specific implementation guidelines that can adapt to different organizational needs while maintaining consistency in benefit assessment. This includes establishing mechanisms for long-term impact assessment and creating standardized measurement tools that can adapt to specific use cases while maintaining comparability across implementations. Furthermore, the protocol should incorporate mechanisms for ongoing assessment and adjustment to account for evolving technology landscapes and emerging stakeholder needs.

5.2 Limitations

Our benefit assessment protocol faces several key limitations identified through the pilot interviews. A fundamental challenge lies in establishing direct one-to-one mappings between risks and benefits. While the protocol provides a structured approach to identifying both benefits and risks, participants highlighted that benefits often have multiple associated risks, and conversely, risks can impact multiple benefit areas. For example, the efficiency gains in automated report generation may simultaneously carry risks of job displacement and data privacy concerns. As next steps, our protocol could be designed with a more comprehensive literature review surrounding AI benefits to develop a benefit theme taxonomy, and the protocol could be updated by conducting more interviews and iteratively incorporating feedback from participants into the protocol.

REFERENCES

[1] Paul Anand, Graham Hunter, Ian Carter, Keith Dowding, Francesco Guala, and Martin Van Hees. 2009. The development of capability indicators. *Journal of Human Development and Capabilities* 10, 1 (2009), 125–152.

[2] Jacqui Ayling and Adriane Chapman. 2022. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics* 2, 3 (2022), 405–429.

[3] Microsoft Corporation. 2022. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf

[4] EqualAI. [n. d.]. Equalai Algorithmic Impact Assessment (AIA). https://www.equalai.org/aia/#:~:text=The%20EqualAI%20Algorithmic%20Impact%20Assessment,in%20the%20recent%20NIST%20publications.&text=Our%20example%20EqualAI%20AIA%20is,and%20does%20not%20collect%20information.

[5] Baruch Fischhoff. 2015. The realities of risk-cost-benefit analysis. *Science* 350, 6260 (2015), aaa6516.

[6] Richard Fulton, Diane Fulton, Nate Hayes, and Susan Kaplan. 2024. The Transformation Risk-Benefit Model of Artificial Intelligence: Balancing Risks and Benefits Through Practical Solutions and Use Cases. *arXiv preprint arXiv:2406.11863* (2024).

[7] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.

[8] Alex John London and Hoda Heidari. 2024. Beneficent intelligence: a capability approach to modeling benefit, assistance, and associated moral failures through AI systems. *Minds and Machines* 34, 4 (2024), 41.

[9] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 735–746.

[10] Jimin Mun, Liwei Jiang, Jenny Liang, Inyoung Cheong, Nicole DeCairo, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. 2024. Particip-ai: A democratic surveying framework for anticipating future ai use cases, harms and benefits. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 997–1010.

[11] United Nations. 2024. Human Development Index. https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

[12] Martha C Nussbaum. 2000. *Women and human development: the capabilities approach*. Cambridge University Press.

[13] Government of Canada. 2024. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html#toc2

[14] Government of the Netherlands. 2023. https://www.government.nl/documents/publications/2023/03/02/ai-impact-assessment

[15] UN Global Pulse. 2020. https://www.unglobalpulse.org/document/risks-harms-and-benefits-assessment-tool/

[16] M.J. Sandel. 2010. *Justice: What's the Right Thing to Do?* Farrar, Straus and Giroux. https://books.google.com/books?id=BrdNDG7TTUEC

[17] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. 2020. Principles to practices for responsible AI: closing the gap. *arXiv preprint arXiv:2006.04707* (2020).

[18] Data Science and Public Policy. 2021. https://datasciencepublicpolicy.org/our-work/tools-guides/data-science-project-scoping-guide/

[19] Amartya Sen. 1999. Commodities and capabilities. *OUP Catalogue* (1999).

[20] Somesh Sharma. 2024. Benefits or concerns of AI: A multistakeholder responsibility. *Futures* (2024), 103328.

[21] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[22] Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv preprint arXiv:2408.12622* (2024).

[23] Cass R Sunstein. 2005. Cost-benefit analysis and the environment. *Ethics* 115, 2 (2005), 351–385.

[24] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986* (2023).

A PILOT STUDY INTERVIEW QUESTIONS

We are conducting a study on the usability and utility of a benefit analysis protocol for Al systems. This protocol aims to systematically assess who benefits, how they benefit, the trade-offs involved, potential risks, as well as define appropriate metrics and measures for these factors. We are seeking feedback from experts like you to refine and validate its practical applicability.

We would like to record the interview. The recording will only be used to write the class project report, and will not be shared with anyone outside of the project team. Do you consent to being recorded?

INTRODUCTION AND THOUGHTS ON THIS PROBLEM

Q1. When thinking about assessing the benefits, risks, and trade-offs of an Al system, what approaches or methods come to mind?

Q2. Have you seen or used any frameworks for this in your work?

AL SYSTEM IDENTIFICATION AND DESCRIPTION

Q3. Can you think of an Al system that impacts members of the general public? It could be something you've worked on or are familiar with.

Q4. Can you describe this system for me? What does it do, and how is it used?

STAKEHOLDER IDENTIFICATION

Q5. Who are the stakeholders affected by this system? (Consider direct users, decision-makers, external groups, marginalized communities, and any others impacted by its deployment)

LISTING BENEFITS, RISKS AND COSTS FOR EACH STAKEHOLDER

Q6. What are the benefits, risks, and costs experienced by each stakeholder group you identified? Can you give specific examples?

METRICS

Q7. What kind of metrics would you use to quantify the benefits of this system?
- Follow-ups: Would you use quantitative variables? Why or why not?
- Would you use qualitative variables? Why or why not?
- Are there any existing metrics or methods you'd recommend for this purpose?

Q8. How would you measure risks or costs associated with this system?
- Follow-ups: Would you use quantitative variables? Why or why not?
- Would you use qualitative variables? Why or why not?
- Are there existing metrics or methods you'd recommend for this purpose?

USABILITY FEEDBACK ON OUR PROTOCOL:

BENEFIT ASSESSMENT PROTOCOL DRAFT 1

(Show Our Documentation) [Give a few minutes to go through it]

Q9. How does our approach compare to your initial thoughts on how to assess benefits? Are there any steps or considerations you feel are missing? Are there any steps or considerations you feel are unnecessary?

Q10. When considering the stakeholders of an AI system, do you find our categories (internal, external - [voiceless, vested interest, oversight]) sufficient? If not, who else should be considered, and why?

Q11. What are your thoughts on our measures and metrics of scoring benefits, risks, and costs?

- How would you improve the measures and metrics?

PRACTICAL CHALLENGES

Q11. Do you foresee any challenges in applying this protocol to real-world systems?

FEEDBACK

Q12. What do you think are the strengths of this protocol?

Q13. What do you think are the weaknesses or areas that need more work?

Q14. If you were to use this protocol in your own work, what support or resources would you need to apply it effectively?

OPTIONAL ADDITIONAL FEEDBACK

Q15. Is there anything else you'd like to share that we haven't covered?