Northwestern | CENTER FOR **ADVANCING SAFETY OF MACHINE INTELLIGENCE**

AI Safety: A Domain-Focused Approach to Anticipating Harm

INTRODUCTION **2**

PROCESS **2**

DOMAIN LEVEL HARMS AND BENEFITS 3

ENVISIONING HARM SCENARIOS 3

COMMUNICATION REQUIREMENTS AND PLANS 3

TALKS/PANELS **3**

DOMAIN LEVEL HARMS AND BENEFITS **4**

MEDICINE 4

Benefits *5*

Harms *9*

JOURNALISM 14

Benefits *14*

Harms *16*

LAW 20

Benefits *20*

Harms *22*

ENVISIONING HARM SCENARIOS **26**

MEDICINE 27

Generative AI (Tiered Medical Services) *27*

Predictive Analytics (Faulty Training Sets) *28*

Recommendation Systems (Skewed Professional Advancement) *28*

JOURNALISM 29

Generative AI (Generation of False News) *29*

Predictive Analytics (Summarization of News) *30*

Recommendation Systems (Information Specialization for Profits) *30*

LAW 31

Generative AI (Leveling of Reasoning) *31*

Predictive Analytics (Lower Quality of Legal Services) *32*

Recommendation Systems (Digital Twin Replaces) *33*

COMMUNICATION REQUIREMENTS AND PLANS **33**

MEDICINE 34

Existing Methods for Tracking AI Harms *34*

Information Requirements/Suggestions *36*

JOURNALISM 39

Existing Methods for Tracking AI Harms *39*

Information Requirements/Suggestions *41*

LAW 42

Existing Methods for Tracking AI Harms *42*

Information Requirements/Suggestions *45*

OUTCOMES **47**

NEXT STEPS **48**

CASE GATHERING AND CURATION 48

VALIDATION AND ANALYSIS 48

DOMAIN- AND ROLE-CENTRIC DOCUMENT GENERATION 49

Introduction

Artificial Intelligence (AI) now intersects with essentially every aspect of our lives. You can find it in business, media, entertainment, marketing, medicine, and education. Unfortunately, the rollout of the technologies has not included a thoughtful consideration of how it will impact our work, lives, and society. Even with the introduction of regulations in the US and the enactment of rules in the EU, the thinking regarding AI remains vague and unfocused. It lacks a sense of specificity and frames issues in terms of ethical abstractions such as fairness, transparency, and bias that are nearly impossible to operationalize. They sound good but often lack true meaning.

Tied to this lack of specificity is the drive to address the issues of Ethical AI through testing, measurement, and evaluation abstracted away from human impact. The notion is that the problems of AI can be explored, identified, and addressed without considering how they are deployed and used in practice. In a world in which our cars remind us to use our seatbelts because of the human tendency to "forget" to do things that are inconvenient, the idea that we are going to understand the impact of AI without considering human reasoning and tendencies is hard to comprehend.

This focus on problems that are abstractions of the real issues which ignore human interaction has resulted in a situation in which technologists and application providers use the excuse that we cannot hope to predict the real problems, and so we won't even try.

To address these issues, we convened a meeting with the Rockefeller Foundation aimed at exploring the different needs and requirements of different industries associated with the benefits and harms of AI. This convening was aimed at considering AI and its impact in a somewhat different way. Our goals were to examine AI from the perspective of human interaction and harm, establish methods for determining the sources of those harms, discover approaches on how to mitigate them in the context of current systems, and to discuss how to avoid harms when developing new systems.

Process

Our approach to exploring industry or domain level issues of harm and benefits had three components.

First, we brought together practitioners in three crucial fields: Medicine, Law, and Journalism; to outline the concerns within their fields regarding AI and their own mechanisms for dealing with how to structure responses within their areas. The goal was to leverage the goals and values of the different fields and develop specific models of harms and safety that can be used to inform their respective ethical stances.

Second, we brought in speakers and panelists to frame certain issues (what constitutes harm, existing mechanism, domain level envisionment strategies) to provide a framework and shared language for our participants.

Third, we explored a suite of approaches to uncovering areas of possible harm, mining existing notions of harm and benefit, envisioning plausible problems, and exploring communication methods for each domain.

The result of this process was a suite of domain level harms and benefits, a validation of a set of approaches (in particular, the envisioning exercise), and the outline of a plan for collecting and communicating harms for each domain.

Domain Level Harms and Benefits

To explore the benefits and harms associated with AI systems in various industries, participants engaged in an exercise to identify, categorize, and map these impacts, ultimately framing opportunities for improvement. This process resulted in a prioritized list of benefits and harms and actionable "How Might We?" statements to guide responsible AI development and deployment.

Envisioning Harm Scenarios

To systematically explore the potential impact of AI technologies on various fields, we used **situation prototyping** and **creative scenario-writing** methods. Participants from medicine, law, and journalism, along with technologists, brainstormed and wrote forward-looking scenarios that included specific settings, characters, and events, focusing on generative AI, predictive analytics, and recommendation engines. Following interactive group sessions and critiques, participants left with a deeper understanding of AI's possible future impacts and the skills to continue exploring these scenarios in their respective fields.

Communication Requirements and Plans

To systematically explore how AI impacts are tracked and understood, we conducted an exercise focusing on key questions about information gathering and reporting in various fields. Participants examined how their fields currently track issues, specifically those related to AI, and discussed potential mechanisms for improving AI issue tracking. This exercise facilitated a comprehensive understanding of existing practices and inspired ideas for new, effective methods of monitoring AI-related problems.

Talks/Panels

Five speakers gave individual talks to help frame the workshop. We also invited experts in law, journalism, and medicine to participate in a virtual panel titled, “What Constitutes Harm: Different fields and metrics.”

Our first speaker was James Guszcza, who discussed the concept of “Hybrid Intelligence,” a sociotechnical approach to applied AI. While machine learning operations often focus on data engineering, development operations, and machine learning, what we typically need is a design-focused field that is grounded in behavioral sciences, computation, and statistics. Hybrid Intelligence involves all these components to design for real-world goals and to foster decision environments that reflect the needs, behaviors, and cognitive capabilities of the human partner.

Ryan Jenkins gave a talk focused on the research he is conducting with CASMI, “Exploring Methods for Impact Quantification.” This approach compiles an ecumenical theory of human wellbeing to generate a Human Impacts Scorecard, which quantifies the impact of machine learning (ML) systems. This theory of human capabilities is versatile in that it works across multiple domains and applications; holistic and exhaustive in that it provides a rich view of modes of human flourishing; and universal for its application across times, cultures, religions, etc.

These two talks preceded the virtual panel, which Guszcza moderated. The four panelists—Paula Goldman, chief ethical officer at Salesforce; Melissa Goldstein, associate professor in health policy and management at George Washington University; Steven Levy, editor at large at Wired; and Daniel W. Linna Jr., senior lecturer at Northwestern Pritzker School of Law—discussed how their respective fields are currently dealing with AI, the risks and harms they are most worried about, and the measures they are currently taking to mitigate and avoid harms. The panel further informed participants about the framing of the workshop, which was aimed at taking a domain-focused approach toward anticipating AI harms.

A crucial component of anticipating AI harms is to assess the impacts on people. Abigal Jacobs, associate professor of information at University of Michigan, explored this issue during her talk, titled “The hidden governance of AI: Sociotechnical aspects of measurement.” She discussed several case studies regarding disparities in data among protected groups, as well as assumptions that all technical systems depend on to shape outcomes. The talk emphasized identifying errors and proxies in measurement to build better systems.

On the second day of the workshop, Reva Schwartz, research scientist for the National Institute of Standards and Technology (NIST), gave a talk about how the federal agency is using a new program to advance sociotechnical testing and evaluation for AI. The program is called ARIA, which stands for Assessing Risks and Impacts of AI. It considers AI beyond the model and assesses systems in context, including what happens when people interact with AI technology in realistic settings under regular use. This gives a broader, more holistic view of the net effects of these technologies.

The final talk of the workshop focused on the AI Incident Database (AIID), which shares information about AI failures so that they are not repeated. Patrick Hall, assistant professor of decision sciences at George Washington University and leading contributor of incident reports to the AIID, discussed the need to track AI harms and near harms. He also demonstrated how the database works and presented its taxonomy.

Domain Level Harms and Benefits

Our first exercise was aimed at exploring the benefits and harms associated with the use of AI systems in our three industries. The primary goals were to identify and categorize these benefits and harms, map their impacts, and frame opportunities to maximize the positive impacts while mitigating the negative ones. The exercise utilized a structured approach, beginning with participants writing down as many benefits and harms as possible, grouping similar ideas, and mapping them based on their perceived impact.

We broke the workshop into domain-focused working groups. The process began with participants individually identifying benefits and harms, which were then collectively grouped and discussed. This was followed by a mapping exercise to assess the relative impact of each identified item. Participants then used the "How Might We?" (HMW) framework to reframe the most impactful benefits and harms into opportunities for improvement and innovation. Each group shared their top HMW statements with others, fostering a collaborative environment for identifying actionable insights.

The outcomes of this exercise included a comprehensive list of benefits and harms related to AI usage, a prioritized map highlighting high-impact areas, and a set of HMW statements providing a foundation for future initiatives. These outcomes serve as valuable inputs for guiding the responsible development and deployment of AI technologies, ensuring that they enhance their respective fields while addressing potential risks.

Medicine

Exploring the impact and potential harms of AI in the field of medicine is crucial due to the profound implications these technologies have on healthcare delivery, patient outcomes, and clinical processes. AI's ability to assist in diagnosing diseases, personalizing treatment plans, and predicting patient outcomes with unparalleled accuracy and efficiency represents a transformative advancement in medical practice. However, this integration is not without significant concerns. Patient privacy and data security are paramount issues, as the vast amounts of sensitive information required for AI systems present attractive targets for cyber threats. Moreover, biases embedded in AI algorithms could lead to unequal treatment, exacerbating existing healthcare disparities. The risk of over-reliance on AI systems may also erode the essential human **elements** of empathy and clinical judgment that are critical ***in-patient*** care. Therefore, a comprehensive examination of AI's impact and potential harms in medicine is essential to ensure these technologies enhance rather than undermine the quality and equity of healthcare. This work aims to address these concerns, providing a balanced understanding to guide the responsible development and deployment of AI in the medical field.

Benefits

Artificial intelligence (AI) is transforming medicine. At our workshop, attendees enumerated the immediate benefits of using AI in health care. The identified benefits include enhancing patient care, streamlining operations, and improving outcomes. AI algorithms detect diseases early, tailor treatment plans, and predict patient outcomes. AI improves medical imaging accuracy, accelerates drug discovery, and supports virtual health assistants and remote monitoring. AI also aids in robotic surgery, administrative efficiency, mental health support, genomic analysis, clinical decision support, telemedicine, and patient data management. Furthermore, AI fosters health trend analysis, accountability, transparency, big data processing, and treatment optimization. By improving decision consistency, prognostic predictions, and physician-patient communication, AI reduces cognitive bias, optimizes medical equipment, predicts disease spread, and enhances diagnosis accuracy, allowing healthcare professionals to focus on higher-level tasks and innovation.

Early Disease Detection: AI algorithms can analyze medical data to detect diseases at their earliest stages, improving patient outcomes. For example, Google's DeepMind has developed AI models that can detect early signs of eye diseases such as diabetic retinopathy and age-related macular degeneration from retinal scans, allowing for earlier intervention and better patient outcomes.

Personalized Treatment Plans: AI can tailor treatments based on individual patient data, leading to more effective therapies. For instance, IBM Watson for Oncology uses AI to analyze patient data and recommend personalized cancer treatment plans, improving the effectiveness of therapies by considering the unique characteristics of each patient.

Predictive Analytics: AI can predict patient outcomes and potential complications, allowing for proactive management. For example, the predictive analytics platform from Health Catalyst uses AI to forecast patient outcomes and identify those at risk of COVID-19 hospitalization, enabling healthcare providers to intervene early and manage patient care more effectively.

Medical Imaging Analysis: AI enhances the accuracy and efficiency of interpreting medical images, aiding in quicker diagnosis. For example, Zebra Medical Vision's AI algorithms analyze medical imaging data to detect conditions such as fractures, diseases, and tumors, improving diagnostic accuracy and speed.

Drug Discovery: AI accelerates the drug development process by predicting how different compounds will behave. For instance, Insilico Medicine uses AI to predict the efficacy and safety of new drug compounds, significantly reducing the time and cost associated with drug discovery and development.

Virtual Health Assistants: AI-powered chatbots and virtual assistants provide medical advice and support to patients. For example, the AI chatbot from Ada Health offers personalized medical advice and symptom assessments, helping patients understand their conditions and seek appropriate care.

Remote Monitoring: AI enables continuous monitoring of patients' health through wearable devices, ensuring timely interventions. For instance, Biofourmis' AI platform uses data from wearable devices to monitor patients' vital signs and detect early signs of deterioration, allowing for timely medical interventions.

Robotic Surgery: AI improves the precision and safety of surgical procedures, reducing recovery times. For example, the da Vinci Surgical System uses AI to enhance the precision of minimally invasive surgeries, leading to faster recovery times and reduced risk of complications for patients.

Administrative Efficiency: AI automates administrative tasks, reducing paperwork and freeing up time for healthcare professionals. For instance, before the company shut down in November 2023, Olive AI automated repetitive administrative tasks such as billing processing and patient scheduling, which allowed healthcare staff to focus more on patient care.

Mental Health Support: AI tools can provide support and early intervention for mental health issues. For example, Woebot's AI-powered chatbot provides cognitive-behavioral therapy and mental health support, helping users manage their mental health more effectively.

Genomic Analysis: AI analyzes genetic information to identify risks and tailor treatments based on genetic makeup. For instance, Deep Genomics uses AI to analyze genomic data and correct the effects of genetic mutations, enabling the development of personalized therapies based on an individual's genetic profile.

Clinical Decision Support: AI provides real-time assistance to doctors, helping them make informed decisions. For example, the AI platform from PathAI provides pathologists with diagnostic insights based on image analysis, aiding in the accurate diagnosis of diseases such as cancer.

Telemedicine: AI enhances telemedicine by providing real-time data analysis and patient monitoring. For instance, the AI-powered platform from TytoCare allows for real-time patient monitoring and data analysis during virtual consultations, improving the quality of remote care.

Patient Data Management: AI ensures accurate and secure handling of patient records and data. For example, Flatiron Health uses AI to configure patient data from electronic health records, ensuring data validation while providing valuable insights for cancer treatment.

Health Trend Analysis: AI identifies trends and patterns in health data, aiding in public health planning. For instance, BlueDot uses AI to analyze global health data and identify trends in disease outbreaks, helping public health officials plan and respond more effectively to emerging health threats.

Accountability: AI can enhance accountability in the medical field by tracking and documenting decision-making processes, ensuring that every action taken is traceable. This can help in identifying errors, improving patient safety, and maintaining high standards of care.

Transparency: The use of AI can increase transparency in healthcare by providing clear, data-driven insights into treatment plans and medical decisions. This helps patients understand their care processes better and builds trust in the healthcare system. For instance, the AI platform from Tempus analyzes clinical and molecular data to provide transparent insights into personalized treatment plans, helping patients and healthcare providers make informed decisions.

Ability to process MORE/BIG data: AI can analyze vast amounts of medical data quickly and efficiently, uncovering patterns and insights that might be missed by human analysis. This ability to handle big data enables more comprehensive research and better patient outcomes. For example, Google DeepMind's AI algorithms process large datasets from various health records to uncover patterns that improve predictive diagnostics and patient outcomes.

Treatment optimization (targeted therapy based on models): AI can optimize treatments by using predictive models to tailor therapies to individual patients, ensuring that treatments are more effective and reducing the likelihood of adverse reactions. For instance, PathAI uses AI to analyze pathology data and optimize cancer treatments by tailoring therapies based on the unique characteristics of each patient’s tumor.

Foster notification of patient status change: AI systems can continuously monitor patient data and alert healthcare providers to significant changes in patient status, allowing for timely interventions and improved patient care. For example, Philips’ IntelliVue Guardian Solution uses AI to monitor patient vital signs and notify caregivers of any significant changes, enabling timely interventions.

Better predictions, better care: AI's predictive capabilities can lead to more accurate forecasts of disease progression and patient outcomes, enabling healthcare providers to plan better and deliver higher quality care. For instance, the AI platform from Aidoc analyzes medical images to predict disease progression and patient outcomes, helping radiologists and clinicians deliver better care.

Consistency of decisions: AI can help standardize medical decisions by providing consistent, evidence-based recommendations, reducing variability in care and ensuring that all patients receive the best possible treatment.

Prognostic prediction (risk of death or hospitalization): AI can predict the likelihood of outcomes such as death or hospitalization based on patient data, helping healthcare providers to identify high-risk patients and implement preventive measures. For example, the AI model developed by Epic Systems can predict the risk of patient deterioration or hospitalization, allowing for early intervention and better management of high-risk patients.

Identify physiologic association: AI can identify complex physiological associations and correlations within patient data, enhancing understanding of diseases and informing more effective treatment strategies. For instance, the AI platform from Zebra Medical Vision identifies physiological associations in imaging data to improve disease diagnosis and treatment planning.

Improve physician empathy, communication with patients: By handling routine tasks and data analysis, AI frees physicians to spend more time interacting with patients, improving communication and empathy in patient care. For example, the AI assistant from Suki helps physicians with documentation and administrative tasks, allowing them to focus more on patient interactions and improving the overall patient experience.

Lessen physician fatigue: AI can reduce physician workload by automating time-consuming tasks, leading to decreased burnout and fatigue, and allowing physicians to focus on patient care. For instance, Nuance's Dragon Medical One uses AI to transcribe and streamline clinical documentation, reducing the administrative burden on physicians and lessening fatigue.

Better care through effective doctor-AI teams: The collaboration between doctors and AI systems can enhance the quality of care, combining human judgment and experience with AI's analytical capabilities to achieve better outcomes. For example, the AI platform from PathAI assists pathologists in diagnosing diseases more accurately and efficiently, resulting in better patient outcomes through collaborative efforts.

Easier access to healthcare: AI can facilitate easier access to healthcare services, particularly in underserved areas, by enabling telemedicine and remote diagnostics, ensuring that more people receive timely medical attention. For instance, eMed (formerly Babylon Health) uses AI to provide remote consultations and diagnostic services, making healthcare more accessible to people in remote and underserved areas.

Removal of human cognitive bias/error: AI can mitigate human cognitive biases and errors in medical decision-making, leading to more accurate diagnoses and treatment plans. For example, Buoy Health's AI-driven symptom checker reduces diagnostic errors by providing unbiased, evidence-based recommendations to patients and healthcare providers.

Medical/shift efficiency: AI can optimize hospital operations and shift schedules, ensuring that resources are used efficiently and that patients receive timely care. For instance, Qventus uses AI to optimize hospital operations, including shift scheduling and patient flow management, improving overall efficiency and patient care.

Optimizing medical equipment: AI can enhance the use of medical equipment by predicting maintenance needs and optimizing operational parameters, reducing downtime and improving patient outcomes. For example, GE Healthcare’s AI-based predictive maintenance system monitors medical equipment to predict maintenance needs and reduce downtime, ensuring that equipment is available when needed.

Prediction of disease spread: AI can model and predict the spread of diseases, aiding in public health planning and the implementation of effective containment strategies. For instance, BlueDot's AI platform predicts the spread of infectious diseases, helping public health officials plan and implement effective containment measures.

Accuracy of diagnosis: AI can improve diagnostic accuracy by analyzing medical images and patient data with high precision, supporting healthcare providers in making more accurate diagnoses. For example, Aidoc's AI platform analyzes radiology images to provide accurate diagnostic insights, supporting radiologists in making more precise diagnoses.

Frees humans to do and think at higher level, evolve: By automating routine tasks, AI allows healthcare professionals to focus on more complex and innovative aspects of patient care, driving the evolution of medical practice. For instance, the AI-driven tools from Tempus automate data analysis, enabling researchers and clinicians to focus on developing new treatments and advancing medical knowledge.

Patient education: AI can enhance patient education by providing personalized information and resources tailored to individual health needs. For example, HealthTap's AI-driven platform offers personalized health advice and educational resources to patients, helping them make informed decisions about their health and adhere to treatment plans more effectively.

Creating new norms: AI has the potential to create new norms in healthcare by standardizing best practices and promoting evidence-based guidelines, leading to more consistent and high-quality care across different settings.

Collective tradeoffs: AI can help healthcare providers and policymakers understand and navigate collective tradeoffs by analyzing large datasets to balance various factors such as cost, quality, and access, ultimately improving decision-making processes. For example, Optum's AI analytics tools help healthcare providers and policymakers analyze large datasets to balance cost, quality, and access, improving decision-making processes and optimizing resource allocation.

Impact on healthcare workforce: The integration of AI in healthcare can significantly impact the workforce by automating routine tasks, allowing healthcare professionals to focus on more complex and value-added activities, and potentially reshaping job roles and responsibilities.

Model that schedules patients off waitlist: AI can develop models that efficiently schedule patients off waitlists, reducing waiting times and ensuring timely access to care for those in need, ultimately improving patient outcomes and satisfaction. For example, Qventus uses AI to create models that efficiently schedule patients off waitlists, reducing waiting times and improving access to care, which leads to better patient outcomes and satisfaction.

Democratization of access to medical knowledge: AI facilitates the democratization of access to medical knowledge by making information readily available to a broader audience, bridging the gap between healthcare providers and patients, and promoting health literacy. For instance, Ada Health's AI-powered app provides medical knowledge and health advice to users worldwide, democratizing access to medical information and promoting health literacy.

Chatbots for improved mental health (e.g., CBT, mindfulness): AI-powered chatbots can offer support for mental health through cognitive-behavioral therapy (CBT) and mindfulness exercises, providing accessible and cost-effective mental health care for individuals who might not have access to traditional therapy. For example, Woebot's AI-driven chatbot provides CBT and mindfulness support to users, offering accessible and cost-effective mental health care for those who might not have access to traditional therapy.

AI-assisted nudges for improved health behaviors: AI can assist in delivering nudges that promote healthier behaviors, such as reminders to take medication, exercise, or attend medical appointments, thereby improving overall health outcomes. For example, Lark Health uses AI to deliver personalized nudges and reminders to users, promoting healthier behaviors such as medication adherence, exercise, and diet, thereby improving overall health outcomes.

Save time: AI can save time for healthcare providers by automating administrative and clinical tasks, streamlining workflows, and enabling more efficient use of resources, allowing providers to dedicate more time to patient care. For instance, Epic's AI-driven EHR system automates administrative and clinical tasks, streamlines workflows, and enables more efficient use of resources, allowing healthcare providers to dedicate more time to patient care.

Harms

At our workshop, participants were also asked to enumerate the harms of using AI in medicine. The discussion brought to light several significant concerns, such as privacy violations, the potential for increased stigma, and lower trust in healthcare. AI's impact on clinician critical thinking, the risk of exacerbating existing health disparities, and issues related to resource allocation were also highlighted. Moreover, the potential for AI to eventually replace expert tasks, create tunnel vision, and foster a shallow understanding of medical contexts were noted. These harms underscore the need for careful implementation and regulation of AI in healthcare to prevent unintended negative consequences.

Privacy violations: AI systems require vast amounts of patient data, raising significant concerns about privacy violations. For example, a breach in an AI system handling sensitive patient information could expose personal health records, compromising patient confidentiality.

Stigma: The use of AI in medicine can lead to stigma if certain conditions or behaviors are flagged more frequently by AI systems. For instance, an AI tool that predicts mental health issues might inadvertently label patients, leading to stigmatization and discrimination.

Lower trust in healthcare: The increasing reliance on AI in medicine can lower trust in healthcare if patients feel that decisions are being made by machines rather than humans. For example, patients might distrust AI-generated treatment plans, fearing that they lack the human touch and nuanced understanding that doctors provide.

Impact on clinician/human critical thinking: Over-reliance on AI can diminish critical thinking skills among clinicians. For instance, if doctors rely too heavily on AI for diagnosis, they might lose essential diagnostic skills over time, compromising their ability to make independent clinical judgments.

Exaggerate existing health differences: AI systems can exacerbate existing health disparities if they are not designed and trained inclusively. For example, an AI tool trained on data from predominantly affluent populations might not perform well for patients from low-income backgrounds, leading to unequal treatment outcomes.

Resource allocation (who gets money or resources): AI can influence how resources are allocated in healthcare, potentially leading to inequities. For instance, AI might prioritize certain treatments or patient groups over others, leading to unfair distribution of medical resources.

Eventually replace/automate many if not all expert tasks: The automation of expert tasks by AI could eventually replace many roles currently filled by medical professionals. For example, AI systems might take over diagnostic and treatment planning tasks, reducing the need for human specialists and potentially leading to job losses.

Tunnel vision (limited imagination): AI systems might create tunnel vision by focusing too narrowly on specific data points or outcomes. For instance, an AI tool that prioritizes efficiency might overlook holistic aspects of patient care, resulting in limited treatment approaches.

Shallow understanding: AI systems might offer a shallow understanding of complex medical conditions due to their reliance on data patterns rather than deep clinical knowledge. For example, an AI diagnostic tool might identify a disease based on symptoms but fail to consider the broader context of the patient's health.

Lack of accountability: Determining accountability for errors made by AI systems can be challenging. For instance, if an AI system misdiagnoses a patient, it can be difficult to determine whether the responsibility lies with the AI developers, healthcare providers, or the healthcare institution.

Forced homogeneity: AI can lead to forced homogeneity in treatment approaches by standardizing medical decisions. For example, an AI system might recommend the same treatment protocol for all patients with a particular condition, ignoring individual variations and preferences.

Lack of choice/notice: Patients may not always be informed or given a choice about the use of AI in their care. For instance, a hospital might implement an AI-based diagnostic tool without adequately informing patients or obtaining their consent, potentially undermining patient autonomy.

Trending data imbalance, leading to bias: AI systems can reinforce biases if they are trained on unbalanced data. For example, an AI tool trained on data from a predominantly white population might underperform for minority groups, leading to biased treatment recommendations.

Bias: AI algorithms can perpetuate and amplify existing biases in healthcare data. For example, an AI system used for predicting patient outcomes might consistently perform worse for certain demographic groups, leading to unequal treatment.

Machine-driven errors: AI systems can make errors, sometimes with severe consequences. For instance, an AI tool might misinterpret medical images, leading to incorrect diagnoses and inappropriate treatments.

Mental/emotional/behavioral/physical harm (medical errors): Errors made by AI systems can cause significant harm to patients. For example, a wrong dosage recommendation from an AI system could lead to adverse drug reactions, causing physical harm to the patient.

Blind trust and misinformation: Over-reliance on AI systems can lead to blind trust and the spread of misinformation. For instance, doctors might accept AI recommendations without question, even if the system's data is flawed or the context is misunderstood.

Decreased clinical advancement (e.g., death predictions): Over-reliance on AI predictions might hinder clinical advancements. For example, an AI system predicting high mortality might discourage further treatment efforts, limiting opportunities for clinical innovation and learning.

Lack of empathy/trust/caring/love: AI systems lack the human qualities of empathy, trust, and caring, which are crucial in healthcare. For instance, patients might feel less cared for if their interactions are primarily with AI systems rather than human caregivers.

Diminished physician engagement: AI can reduce the engagement and involvement of physicians in patient care. For example, doctors might become less involved in the diagnostic process if AI systems handle most of the analysis, potentially leading to decreased job satisfaction.

Impact on environment: The development and operation of AI systems can have environmental impacts. For example, the energy consumption of large AI data centers can contribute to carbon emissions and environmental degradation.

Incorrect transcription of recording of patient visits: AI transcription systems might misinterpret recordings, leading to errors in patient records. For instance, an AI tool might incorrectly transcribe a doctor's verbal notes, resulting in inaccurate medical records and potential treatment errors.

Automating patient follow-up communication (Q&A): AI systems used for automating patient follow-up communication might lack the nuance of human interaction. For example, an AI chatbot might not fully address a patient's concerns during follow-up, leading to dissatisfaction and potential miscommunication.

Lower patient-physician relationship/interaction: Increased use of AI might reduce direct interactions between patients and physicians. For instance, patients might feel less connected to their healthcare providers if much of their care and communication is mediated by AI systems.

Cost to healthcare system/provider/patient: Implementing AI systems can be costly, impacting the overall cost of healthcare. For example, the high costs associated with developing and maintaining AI technologies might be passed on to patients, increasing their financial burden.

Moral hazard: The use of AI in healthcare can introduce moral hazards, where parties take on more risks because they do not bear the full consequences of their actions. For instance, healthcare providers might become overly reliant on AI systems, neglecting their responsibility to verify and validate AI-generated recommendations.

Bias in AI Algorithms: AI systems can perpetuate and amplify existing biases in healthcare data, leading to unequal treatment of patients. For example, an AI system used for diagnosing skin conditions might perform poorly on darker skin tones if it was trained predominantly on images of lighter skin tones, leading to misdiagnosis and unequal treatment.

Privacy Concerns: The use of AI in medicine requires large amounts of patient data, raising concerns about data security and patient privacy. For instance, a breach in an AI system handling sensitive patient information could lead to unauthorized access to personal health records, compromising patient confidentiality.

Diagnostic Errors: AI systems may make diagnostic errors due to incorrect or incomplete data, leading to inappropriate treatments. For example, if an AI diagnostic tool is fed inaccurate patient data, it might misdiagnose a condition, resulting in a patient receiving the wrong treatment.

Lack of Transparency: AI algorithms can be complex and opaque, making it difficult for healthcare providers to understand and trust their decisions. For instance, a black-box AI model used to predict patient outcomes might offer no explanation for its predictions, leading to mistrust among healthcare providers who rely on clear reasoning for clinical decisions.

Job Displacement: The automation of certain tasks in healthcare could lead to job losses for medical professionals and support staff. For example, the implementation of AI systems for administrative tasks might reduce the need for clerical staff, potentially leading to job losses in the healthcare sector.

Over-reliance on Technology: Excessive dependence on AI systems could reduce the clinical skills of healthcare providers. For instance, if doctors rely too heavily on AI for diagnosis, they might lose critical diagnostic skills over time, potentially compromising their ability to make independent clinical judgments.

Data Quality Issues: AI systems require high-quality data, and poor data quality can lead to incorrect predictions and treatments. For example, an AI system trained on incomplete or inaccurate patient data might produce unreliable diagnostic or treatment recommendations.

Regulatory Challenges: The rapid development of AI technologies outpaces regulatory frameworks, leading to potential gaps in oversight and safety. For instance, new AI diagnostic tools might be deployed before adequate regulatory standards are in place, posing risks to patient safety.

Cost of Implementation: The high cost of developing and implementing AI systems can be a barrier for many healthcare institutions. For example, small clinics may struggle to afford the initial investment and ongoing maintenance costs of advanced AI technologies.

Ethical Concerns: The use of AI in medicine raises ethical questions about decision-making, consent, and accountability. For instance, decisions made by AI in patient care might lack the nuanced ethical considerations that human doctors provide, leading to potential ethical dilemmas.

Inequitable Access: Access to advanced AI technologies may be limited to well-funded healthcare institutions, exacerbating health disparities. For example, hospitals in low-income areas might not be able to afford state-of-the-art AI diagnostic tools, resulting in unequal access to high-quality care.

Liability Issues: Determining liability for errors made by AI systems can be legally complex and contentious. For instance, if an AI system misdiagnoses a patient, it can be difficult to determine whether the responsibility lies with the AI developers, healthcare providers, or the healthcare institution.

Data Ownership: Issues around who owns and controls patient data can complicate the use of AI in healthcare. For example, conflicts may arise between patients, healthcare providers, and AI companies over the ownership and control of health data used by AI systems.

Human-AI Interaction: Poorly designed AI systems can lead to frustration and errors in their interaction with healthcare providers. For instance, an unintuitive AI interface might cause healthcare providers to make mistakes or waste time trying to navigate the system.

Algorithmic Misuse: AI algorithms could be misused intentionally, leading to harm or exploitation of patients. For example, an AI system could be manipulated to recommend unnecessary procedures for financial gain, compromising patient care.

Interoperability Problems: AI systems may face challenges in integrating with existing healthcare technologies and databases. For instance, a new AI diagnostic tool might not be compatible with a hospital's electronic health record system, leading to inefficiencies and data silos.

Patient Mistrust: Concerns about AI in healthcare can lead to mistrust and reluctance among patients to accept AI-driven treatments. For example, patients may be wary of AI recommendations if they feel that these systems are not transparent or trustworthy.

Skill Gap: There may be a lack of adequately trained professionals to develop, implement, and manage AI systems in healthcare. For instance, healthcare institutions may struggle to find staff with the necessary expertise to maintain and troubleshoot advanced AI technologies.

False Security: Overconfidence in AI predictions can lead to a false sense of security, potentially overlooking critical issues. For example, doctors might become too reliant on AI diagnostic tools, neglecting to verify results with their clinical judgment, which could result in missed or incorrect diagnoses.

Complexity of AI Systems: The complexity of AI systems can make maintenance and troubleshooting difficult, leading to potential system failures. For instance, a sophisticated AI system might require specialized knowledge to fix issues, and a lack of skilled personnel could lead to prolonged downtimes and compromised patient care.

These potential harms highlight the need for careful consideration and management of AI technologies in medicine to ensure they benefit patients and healthcare providers without introducing new risks.

Journalism

Exploring the impact and potential harms of AI in the field of journalism is crucial due to the significant implications these technologies have on news reporting, information dissemination, and the integrity of media. AI's ability to assist in automating content creation, curating personalized news feeds, and analyzing vast amounts of data to uncover trends and insights represents a transformative advancement in the journalism industry. However, this integration is not without considerable concerns. Issues of misinformation and fake news are paramount, as AI systems could inadvertently generate or spread false information, undermining public trust in the media. Additionally, biases embedded in AI algorithms could lead to unequal representation and coverage, perpetuating existing societal biases. The risk of over-reliance on AI systems may also erode the critical human elements of journalistic judgment, ethics, and the nuanced understanding required for investigative reporting. Therefore, a comprehensive examination of AI's impact and potential harms in journalism is essential to ensure these technologies enhance rather than undermine the quality, accuracy, and fairness of news reporting. This work aims to address these concerns, providing a balanced understanding to guide the responsible development and deployment of AI in the journalism field.

Benefits

Artificial Intelligence (AI) is revolutionizing the field of journalism, offering a range of benefits that enhance the way news is gathered, processed, and delivered. The integration of AI into journalism holds the promise of more personalized, efficient, and accurate news delivery, thereby transforming the media landscape.

At our workshop, attendees were asked to enumerate the immediate benefits of using AI in journalism. The identified benefits include increased efficiency, enhanced accuracy, personalized content delivery, improved investigative capabilities, real-time data analysis, cost reduction, uncovering hidden patterns and trends, personalized news and targeted feeds, AI structuring news with sophisticated data analysis, prosocial ranking, wide reach to different ethnicities, easier access to publication archives, and assistance with drafting and collecting ideas.

Personalized News & Targeted Feeds: Improved customization of news delivery allows users to **receive** content tailored to their interests and preferences. For instance, The New York Times app uses AI to analyze a user’s reading habits and preferences to deliver a personalized news feed, ensuring that the user receives stories that are most relevant to them.

AI Structuring News: Sophisticated Data Analysis: Enhanced data analysis capabilities provide deeper insights and more relevant information to users. For example, Reuters uses AI to analyze complex data sets from financial reports to extract key points and trends, making it easier for journalists to create comprehensive stories.

Prosocial Ranking: AI-driven ranking systems prioritize content based on its potential to promote positive social outcomes and encourage civic engagement. For instance, a local news website might use an AI algorithm to rank articles about community service initiatives higher, thus encouraging readers to engage in civic activities.

Reach: The use of translation to reach different ethnicities who communicate in different languages expands the audience base. For example, BBC News uses AI-powered translation services to provide content in multiple languages, making it accessible to non-English speaking communities in the UK.

Search and Curation: AI enables easier access to publications' archives, allowing for more efficient research.

Brainstorming: AI assists with easy drafting and collecting ideas. For instance, a journalist can use an AI tool to generate topic suggestions and outlines based on trending news topics and keywords.

Enhanced News Gathering: AI tools quickly gather and analyze information from diverse sources, helping journalists stay updated with breaking news. For example, Associated Press uses AI to scan social media platforms and news sites to provide real-time updates on developing stories.

Automated Reporting: AI can generate news articles on specific topics, such as sports scores or financial reports, freeing up journalists for more in-depth reporting.

Fact-Checking: AI algorithms verify information and detect false claims, improving the accuracy and credibility of news reports. For example, PolitiFact uses an AI fact-checking tool to cross-reference statements in political speeches with a database of verified facts to identify inaccuracies.

Personalized Content: AI tailors news content to individual readers' preferences, enhancing user engagement and satisfaction. For instance, Medium uses AI to recommend articles based on a user's previous reading history and interests.

Data Analysis: AI analyzes large datasets to uncover trends and insights, enabling journalists to produce more data-driven stories. For example, FiveThirtyEight uses AI to process election data to highlight voting patterns and demographic shifts.

Natural Language Processing: AI transcribes interviews and converts audio and video content into text, making it easier for journalists to handle multimedia content. For instance, AI at CNN automatically transcribes recorded interviews, saving journalists time on manual transcription.

Audience Engagement: AI analyzes audience behavior and feedback, helping news organizations understand their audience better and create more relevant content.

Content Curation: AI recommends related articles and multimedia content, enriching the reader's experience and keeping them on the site longer. For example, after reading a news article, AI can suggest related stories, videos, and infographics to the reader.

Translation Services: AI provides real-time translation of news articles, making content accessible to a global audience. For instance, a breaking news story can be instantly translated into multiple languages, reaching a broader audience.

Social Media Monitoring: AI tracks trends and public sentiment on social media, providing journalists with insights into what topics are resonating with the audience.

Augmented Reality: AI-powered AR creates immersive news experiences, allowing readers to explore news stories in a more interactive way. For instance, an AR application by The New York Times can provide a virtual tour of a disaster site, offering a deeper understanding of the event.

Predictive Analytics: AI predicts trending topics and upcoming news events, helping journalists plan their coverage in advance. For example, Google AI can predict long-term climate trends and weather, expediting warnings to the public about potential dangers in the forecast.

Image and Video Recognition: AI analyzes and categorizes visual content, making it easier for journalists to find and use relevant images and videos. For instance, AI at Getty Images tags and organizes thousands of photos from a major event, streamlining the editorial process.

Interactive Storytelling: AI helps create interactive news stories, combining text, images, and multimedia elements for a more engaging reader experience. For example, an interactive news story on The Washington Post can include embedded videos, infographics, and interactive maps.

Workflow Automation: AI automates routine tasks, such as scheduling posts and managing content, allowing journalists to focus on more creative work. For instance, AI tools can schedule social media posts to be published at optimal times, ensuring maximum audience reach.

Sentiment Analysis: AI analyzes public sentiment towards news topics, providing insights into audience opinions and reactions. For example, sentiment analysis at BBC News can reveal how readers feel about a controversial issue, helping journalists gauge public opinion.

Enhanced Search: AI improves search functionality on news websites, helping readers find relevant content more efficiently. For instance, an AI-powered search engine on a news website can provide more accurate and relevant search results based on user queries.

Voice Assistants: AI-powered voice assistants provide news updates and answer questions, offering a hands-free way for audiences to stay informed. For example, a voice assistant from NPR can read out the latest news headlines while the user is driving.

Content Moderation: AI monitors and moderates user-generated content, ensuring that discussions remain respectful and on-topic. For instance, AI at Reddit can detect and filter out offensive comments in online forums and comment sections.

Ethical Journalism: AI helps identify and address ethical issues in reporting, promoting fairness and integrity in journalism.

By leveraging these AI capabilities, journalism can become more efficient, accurate, and engaging, ultimately enhancing the quality of news that reaches the public.

Harms

Artificial Intelligence (AI) has the potential to significantly transform the field of journalism, but it also presents numerous potential harms that could negatively impact the industry and society.

During our workshop, attendees discussed the immediate harms of using AI in journalism. The identified harms include potential biases in AI algorithms, reduction in human jobs, spread of misinformation, loss of editorial control, erosion of privacy, dependency on technology, threat to journalistic integrity, high implementation costs, lack of transparency, ethical concerns, reinforcement of existing inequalities, elimination of journalists’ jobs, loss of trust in authenticity, loss of local news outlets, AI threat to journalists' work, lack of industry standards around AI use, lack of disclosure about AI use in reporting, misinformation or persuasion at scale, bad actors using AI to reproduce likenesses of journalists, journalists acting more like AI, paradox of reusability harming AI, predatory licensing deals, and social interaction bots.

Loss of Trust in Authenticity: The rise of AI-generated content may erode public trust in the authenticity and accuracy of news articles. Persuasive-enough deepfakes could become prevalent, collapsing public trust and leading to a dangerous erosion of democracy. For example, a deepfake video of a prominent political figure making controversial statements could be widely circulated on social media, causing public outrage and undermining trust in the media and the democratic process.

Elimination of Journalists' Jobs: Automation and AI could lead to significant job losses within the journalism industry. For example, the biggest-selling newspaper in Europe, Germany's Bild tabloid, cut about 200 jobs in 2023, replacing a range of editorial roles with AI. This displacement can result in fewer opportunities for budding journalists to gain experience and enter the profession.

Readers' Reset: Loss of Local News Outlets: Local news outlets may struggle to compete with AI-driven news platforms, leading to a decline in local news coverage. For example, a small-town newspaper in rural America might be forced to shut down because it cannot match the speed and cost-efficiency of AI-generated news from larger platforms, leaving the community without a dedicated source of local news and important local issues unreported.

AI Threat to Journalists' Work: AI could remove the need for journalists altogether. For example, a news website could use AI to generate all of its content, from news reports to opinion pieces, eliminating the need for human journalists and reducing the diversity of voices in the media. This could lead to homogenized content and a lack of nuanced perspectives on complex issues.

Erosion of Skills: AI could replace many functions traditionally performed by journalists, from data gathering to content creation. For example, journalists at a major news network might rely on AI to analyze data and write reports, leading to a gradual decline in their investigative and writing skills. Over time, this reliance on AI could erode the foundational skills that define quality journalism.

Erosion of Standards: The lack of industry standards around the appropriate and acceptable use of AI in journalism could lead to inconsistent and unreliable practices. For example, without clear guidelines, news organizations might use AI to cut corners in reporting, resulting in lower-quality journalism. This can result in poorly researched articles and the spread of misinformation.

Lack of Transparency: A lack of disclosure about the use of AI in reporting and publishing can lead to a loss of trust among readers. For example, a news article generated by AI might not disclose its origin, misleading readers into believing it was written by a human journalist. This lack of transparency can create a disconnect between news organizations and their audiences.

Misinformation or Persuasion at Scale: AI can be used to generate persuasive but false content, spreading misinformation widely. For example, AI-generated fake news stories could be distributed on social media to influence public opinion on critical issues such as elections or public health. This misuse of AI can manipulate public perception and undermine democratic processes.

Deepfakes in the News: Bad actors using AI to reproduce the likeness of journalists can mislead audiences. For example, a deepfake video of a respected journalist endorsing a false narrative could be used to manipulate public perception and trust. This can damage the reputation of the journalist and the news organization they represent.

Normalization/Standardization of Skills: Journalists might begin to act more like AI, following standardized methods that lack creativity and critical thinking. For example, newsrooms might adopt rigid AI-driven templates for reporting, stifling individual journalistic flair and investigative depth. This can lead to a homogenized media landscape with little room for innovative storytelling.

Paradox of Reusability: Harm to AI from degrading primary sources of data with low information content. For example, if AI systems are trained on a growing body of AI-generated content, the quality and reliability of information could decline over time, leading to a feedback loop of misinformation. This can degrade the overall quality of information available to the public.

Social Interaction Bots: The use of AI bots for social interaction and news dissemination could lead to misinformation and reduced quality of human interaction. For example, AI chatbots might engage with readers on news websites, spreading false information and diminishing meaningful human engagement. This can erode the quality of discourse in online communities.

Privacy Concerns: The use of AI in journalism involves handling large amounts of data, raising concerns about the privacy and security of sources and readers. For example, an AI tool used to gather information from social media might inadvertently expose private details about individuals. This can lead to breaches of confidentiality and the erosion of trust in news organizations.

Loss of Editorial Control: Over-reliance on AI for content curation and generation may lead to a loss of editorial oversight and quality control. For example, an AI system might prioritize sensational stories to drive clicks, compromising the journalistic integrity of the news outlet. This can lead to a decline in the quality of journalism and a focus on clickbait content.

Lack of Transparency: AI algorithms can be opaque, making it difficult for readers and journalists to understand how decisions are made. For example, readers might not know why certain news stories are promoted over others, leading to distrust in the news platform. This lack of transparency can undermine trust in the media.

Decreased Accountability: The use of AI can blur the lines of accountability in journalism, making it harder to attribute responsibility for errors or bias. For example, if an AI-generated article contains inaccuracies, it might be unclear who is responsible for the mistake—the journalists, the AI developers, or the news organization. This can complicate efforts to address and correct errors.

Erosion of Trust: If AI-generated content is perceived as less trustworthy or credible, it can erode public trust in news organizations. For example, widespread use of AI in news production might lead to skepticism among readers about the authenticity of the content they consume. This can damage the credibility of news organizations.

Manipulation and Propaganda: AI can be used to create deepfakes and other manipulative content, spreading propaganda and misleading information. For example, a deepfake video showing a public figure making false statements could be used to sway public opinion. This can undermine democratic processes and trust in public figures.

Quality of Content: AI-generated articles may lack the depth, nuance, and critical analysis that human journalists provide. For example, an AI-written news story might cover the facts but miss the context and insights that a human journalist would include. This can result in superficial coverage of complex issues.

Ethical Concerns: The use of AI in journalism raises ethical questions about consent, bias, and the potential for misuse. For example, AI might be used to gather information in ways that violate privacy or ethical standards. This can lead to ethical dilemmas and the erosion of journalistic integrity.

Reduced Diversity of Perspectives: AI algorithms may prioritize popular or mainstream viewpoints, reducing the diversity of perspectives in news coverage. For example, niche or minority voices might be underrepresented in AI-curated news feeds. This can result in a less diverse and inclusive media landscape.

Algorithmic Censorship: AI-driven content moderation can inadvertently censor legitimate news stories and stifle free expression. For example, an AI system might incorrectly flag and remove articles discussing sensitive but important topics. This can limit the free flow of information and the discussion of critical issues.

Economic Inequality: Access to advanced AI technologies may be limited to large, well-funded news organizations, exacerbating economic disparities in the industry. For example, smaller news outlets might struggle to compete with AI-powered giants, leading to market consolidation. This can reduce the diversity of media voices and perspectives.

Loss of Human Touch: The automation of news generation and reporting can lead to a loss of the human touch and empathy in storytelling. For example, AI-generated content might lack the personal anecdotes and emotional resonance that human writers bring to their stories. This can result in less engaging and relatable news stories.

Dependency on Technology: Over-reliance on AI can make news organizations vulnerable to technical failures and cyber-attacks. For example, a major news website might experience a significant outage if its AI systems fail or are hacked. This can disrupt the dissemination of news and erode public trust.

Intellectual Property Issues: The use of AI to generate content can raise questions about intellectual property rights and ownership. For example, there might be disputes over who owns the rights to AI-generated articles or multimedia content. This can lead to legal challenges and complications.

Interference with Journalistic Ethics: The integration of AI in journalism can create conflicts with traditional journalistic ethics and standards. For example, AI might prioritize sensationalism over accuracy, leading to ethical dilemmas for journalists. This can undermine the integrity and credibility of journalism.

Misinterpretation of Data: AI algorithms can misinterpret data or context, leading to inaccurate or misleading news stories. For example, an AI system might misinterpret statistical data, resulting in a news story that misleads readers about the significance of certain trends. This can contribute to public misunderstanding of important issues.

Monetization Challenges: The reliance on AI-generated content can complicate efforts to monetize journalism and sustain quality reporting. For example, news organizations might struggle to find revenue models that support the high costs of AI development and maintenance while maintaining journalistic quality. This can threaten the financial viability of news organizations.

By addressing these potential harms, the journalism industry can better navigate the challenges posed by AI and ensure that the technology is used ethically and responsibly.

Law

Exploring the impact and potential harms of AI in the field of law is crucial due to the significant implications these technologies have on legal processes, judicial outcomes, and the practice of law. AI's ability to assist in legal research, case analysis, and predicting case outcomes with unprecedented efficiency and accuracy represents a transformative advancement in the legal profession. However, this integration is not without considerable concerns. Issues of privacy and data security are paramount, as the vast amounts of sensitive information required for AI systems present attractive targets for cyber threats. Additionally, biases embedded in AI algorithms could lead to unequal treatment, perpetuating existing legal inequalities and injustices. The risk of over-reliance on AI systems may also erode the critical human elements of judgment, discretion, and ethical considerations that are essential in legal practice. Therefore, a comprehensive examination of AI's impact and potential harms in law is essential to ensure these technologies enhance rather than undermine the quality and fairness of legal services. This work aims to address these concerns, providing a balanced understanding to guide the responsible development and deployment of AI in the legal field.

Benefits

Participants were asked to enumerate the benefits of using AI in the legal field. The consensus highlighted several key advantages, including improved efficiency in legal research, enhanced contract review processes, predictive analytics for case outcomes, and automation of document creation. Other notable benefits include improved e-discovery, compliance monitoring, and intellectual property management. AI's role in legal analytics, case management, fraud detection, and virtual legal assistance also received significant attention. Each of these benefits carries the potential to transform the legal profession, making it more efficient, accurate, and accessible.

Legal Research: AI streamlines legal research by quickly analyzing vast amounts of legal texts, precedents, and case law, saving time and improving accuracy. For example, platforms like ROSS Intelligence use AI to analyze legal documents and provide relevant case law in seconds, enabling lawyers to find pertinent precedents faster and with greater precision.

Contract Review: AI automates the review and analysis of contracts, identifying key terms, potential risks, and inconsistencies. For instance, Kira Systems employs machine learning to identify and extract key clauses from contracts, highlighting potential issues such as non-standard clauses or missing critical terms, which senior legal counsel can then review.

Predictive Analytics: AI predicts case outcomes based on historical data, helping lawyers strategize and make informed decisions. For example, Lex Machina uses predictive analytics to forecast the outcomes of patent litigation by analyzing past cases with similar circumstances, helping lawyers develop better strategies and negotiate favorable settlements.

Document Automation: AI automates the creation of legal documents, reducing manual labor and minimizing errors. For instance, LegalMation leverages AI to generate initial drafts of pleadings and discovery documents for litigators, saving time and ensuring consistency across documents for large corporations.

E-Discovery: AI enhances the e-discovery process by efficiently identifying relevant documents and data during litigation. For example, Relativity's e-discovery platform utilizes AI to sift through millions of emails and documents, identifying those relevant to a case, streamlining the discovery process and reducing costs.

Compliance Monitoring: AI continuously monitors regulatory changes and ensures compliance, reducing the risk of legal issues. For instance, Compliance.ai uses AI to track new financial regulations and automatically update compliance protocols for banks and financial institutions, helping them stay ahead of regulatory requirements and avoid costly fines.

Intellectual Property Management: AI manages IP portfolios, monitors for infringement, and assists in patent analysis and strategy. For example, Anaqua's IP management software uses AI to analyze patent applications in the biotechnology sector, identifying potential overlaps and infringements, helping firms protect their intellectual property more effectively.

Legal Analytics: AI provides insights and trends from legal data, aiding in decision-making and strategic planning. For instance, Bloomberg Law's AI-powered analytics tools help legal professionals analyze litigation trends in the financial services industry, allowing firms to adjust their strategies accordingly and anticipate future legal challenges.

Case Management: AI streamlines case management by organizing and tracking case details, deadlines, and documents. For example, Clio uses AI to automatically update case statuses and send reminders about upcoming deadlines for busy law firms, ensuring that nothing falls through the cracks and improving overall efficiency.

Fraud Detection: AI detects fraudulent activities in financial transactions and corporate activities, aiding in legal investigations. For instance, Symphony AyasdiAI’s platform analyzes transaction patterns in real-time to identify suspicious activities in a corporate merger, providing critical evidence for fraud investigations and helping to prevent financial crimes.

Virtual Legal Assistants: AI-powered virtual assistants provide legal advice and support to clients, improving accessibility to legal services. For example, DoNotPay's AI-driven chatbot offers legal advice and guides users through the process of filing small claims court cases, making legal help more accessible to those who can't afford traditional services.

Litigation Support: AI assists in trial preparation by organizing evidence, identifying key points, and predicting opposing arguments. For example, Everlaw's AI platform helps lawyers prepare for high-stakes trials by organizing thousands of pieces of evidence and suggesting the most relevant arguments based on historical case data, improving the chances of a successful outcome.

Client Due Diligence: AI automates background checks and due diligence processes, ensuring thorough and efficient client vetting. For instance, Intapp uses AI to quickly compile and analyze data on potential clients in corporate transactions, helping firms make informed decisions about client engagements and reducing the risk of working with high-risk individuals or entities.

Risk Assessment: AI evaluates potential legal risks in various scenarios, helping firms mitigate potential issues.

Billing and Time Tracking: AI automates billing processes and tracks billable hours, improving accuracy and efficiency. For instance, TimeSolv uses AI to automatically track the time lawyers spend on different tasks, generating accurate invoices and reducing administrative burdens, allowing lawyers to focus more on their legal work.

Legal Transcription: AI transcribes legal proceedings and meetings, ensuring accurate and searchable records. For example, Verbit's AI-driven transcription service can transcribe court hearings in real-time during complex litigation cases, providing accurate records that can be easily searched and referenced later, aiding in case preparation and appeals.

Expert Witness Identification: AI identifies and evaluates potential expert witnesses for legal cases, based on their expertise and history. For instance, Expert iQ uses AI to analyze databases of expert witnesses to find the most qualified individuals for medical malpractice cases, based on their previous testimonies and expertise, ensuring that the best experts are selected to support the case.

Dispute Resolution: AI facilitates alternative dispute resolution processes, such as mediation and arbitration, by analyzing case details and suggesting solutions. For example, Online Dispute Resolution's AI-driven platform helps parties in commercial disputes by suggesting potential solutions based on the analysis of similar past cases, facilitating quicker and more amicable resolutions without the need for lengthy court proceedings.

Access to Justice: AI provides legal information and resources to the public, improving access to justice for underserved communities.

Talent Management: AI aids in recruiting and managing legal talent, matching candidates with the right skills to appropriate roles. For example, HireVue uses AI to analyze job applications and identify the best candidates for positions in several industries, including law, ensuring that the most qualified individuals are hired.

These benefits underscore the transformative potential of AI in the legal profession, enhancing efficiency, accuracy, and accessibility across various aspects of legal practice.

Harms

At the workshop, participants were asked to provide a list of the potential harms that could result from the application of AI to the practice of law.

The findings highlighted several significant concerns, including potential biases in AI algorithms, the reduction of human jobs, the spread of misinformation, loss of human judgment, privacy issues, dependency on technology, threats to legal integrity, high implementation costs, lack of transparency, ethical concerns, reinforcement of existing inequalities, and the potential misuse of AI. Additional concerns were raised about the loss of agency, unequal access to efficient tools, overreliance on flawed metrics, and leakage of private information. Each of these potential harms carries serious implications for the legal profession and society.

Potential biases in AI algorithms: Bias can lead to unfair outcomes in legal proceedings. AI systems trained on biased data may perpetuate and amplify these biases, resulting in discriminatory practices. For example, if an AI system used for sentencing recommendations is trained on historical data that contains racial biases, it might recommend harsher sentences for minority groups, perpetuating systemic inequalities in the justice system.

Reduction of human jobs: A significant concern among those in the legal industry is that the automation of legal tasks by AI could lead to job losses. This potential displacement of human workers could have widespread economic and social consequences. For example, paralegals and legal researchers might find their roles diminished or eliminated as AI takes over document review and legal research tasks.

Spread of misinformation: AI-generated content can be particularly damaging in the legal field, where accuracy and reliability are paramount. AI tools might inadvertently generate or propagate incorrect legal information, leading to poor decision-making and unjust outcomes. For example, an AI tool providing legal advice might misinterpret a statute or case law, resulting in incorrect guidance to a lawyer or client.

Loss of human judgment: Another critical harm. AI systems, despite their capabilities, lack the nuanced understanding and ethical considerations that human lawyers bring to the practice. Relying too heavily on AI could lead to decisions that are technically correct but lack the wisdom and empathy of human judgment. For example, an AI system might recommend a course of action based solely on legal precedent without considering the broader social or moral implications, leading to outcomes that are legally sound but ethically questionable.

Privacy issues: There are issues with the vast amounts of personal data required to train and operate AI systems. The legal field deals with sensitive and confidential information, and any breach could have severe repercussions. For example, if an AI system used for case management is hacked, sensitive client information could be exposed, leading to identity theft or other forms of exploitation.

Dependency on technology: Dependence on technology can make the legal system vulnerable to technical failures and cyberattacks, which could disrupt legal processes and access to justice. For example, a law firm relying on AI for document management and research might face significant disruptions if their AI system experiences a malfunction or is targeted by a cyberattack, leading to delays and potential loss of critical information.

Threats to legal integrity: There are issues that arise when AI is used without proper oversight and regulation. The integrity of legal proceedings could be compromised if AI systems are manipulated or used unethically. For example, an unscrupulous actor might manipulate an AI system to alter evidence or influence the outcome of a case, undermining the fairness and reliability of the legal process.

High implementation costs: The cost of AI technology can be prohibitive, especially for smaller law firms and legal aid organizations. This could create a disparity between larger firms that can afford advanced AI tools and smaller entities that cannot, leading to unequal access to technological benefits. For example, a small law firm might struggle to compete with larger firms that use AI to streamline their operations and reduce costs, putting the smaller firm at a disadvantage.

Lack of transparency: Lack of transparency in AI decision-making processes can undermine trust in the legal system. If the criteria and algorithms used by AI systems are not transparent, it can be difficult to understand and challenge their decisions. For example, if an AI system used for bail recommendations does not disclose its decision-making process, it might be impossible for defendants or their lawyers to understand why certain recommendations were made or to argue against them.

Ethical concerns: The use of AI in law include issues of accountability, consent, and the potential for misuse. The ethical implications of replacing human judgment with AI need careful consideration to avoid unintended consequences. For example, using AI to predict recidivism rates could lead to decisions that are perceived as unjust or discriminatory if the AI system's predictions are not fully understood or if it relies on biased data.

Reinforcement of existing inequalities: The use of machine learning is a risk when AI systems are not designed and implemented with fairness and inclusivity in mind. AI can exacerbate social and economic disparities if not carefully managed. For example, an AI system that only considers data from well-resourced areas might overlook the needs and circumstances of underrepresented or disadvantaged communities, perpetuating inequality in legal outcomes.

Potential misuse of AI: The legal field includes scenarios where AI is used for unethical purposes, such as surveillance, coercion, or manipulation. For example, an AI system could be used to monitor and influence jurors, compromising the integrity of the trial process and undermining the principles of justice.

Loss of agency/autonomy: Loss of control for lawyers, employees, clients, and civilians is another concern. AI could make decisions or recommendations that limit the ability of individuals to act independently or make informed choices. For example, clients might find their legal options constrained by AI-generated advice that does not fully account for their unique circumstances or preferences.

Unequal access to tools: Given the costs, there is a potential for harm where only well-resourced law firms or clients can afford advanced AI technologies, exacerbating existing inequalities. For example, large law firms might use AI to expedite case preparation, while smaller firms without access to such tools fall behind, leading to a competitive imbalance.

Overreliance on flawed metrics: AI systems prioritizing quantitative measures over qualitative judgment can lead to potentially skewed or incomplete evaluations of legal matters. For example, an AI system might emphasize the number of cases closed rather than the quality of outcomes, incentivizing speed over thoroughness and justice.

Leakage of private information: As in other fields, there is a significant risk as AI systems handle vast amounts of sensitive data, and any breach could have severe consequences for clients and legal professionals. For example, an AI tool used for legal research might inadvertently expose confidential client information to unauthorized parties due to a security vulnerability.

Distracting and wasting resources: Early commitment is a concern if AI systems require significant investment and maintenance without delivering commensurate benefits, diverting resources from other critical areas. For example, a law firm might invest heavily in an AI system that fails to improve efficiency or accuracy, leading to wasted financial and human resources.

Lack of availability of AI tools: Depending on costs, these tools might only be available to the wealthy, which could exacerbate disparities in access to legal services and resources. For example, high-cost AI tools might be used by large corporations to gain an advantage in litigation, while individuals and smaller entities lack access to similar resources, resulting in unequal legal representation.

Propagating inaccurate legal information: AI systems may disseminate incorrect or misleading information, leading to varied levels of legal understanding and application. For example, an AI-generated legal document might contain errors or misinterpretations, causing clients or lawyers to make decisions based on faulty information.

Difficulty in vetting outcomes: Without external validation, AI decisions can undermine trust in the legal process, as it may be challenging to verify the accuracy and fairness of AI-generated results. For example, if an AI system recommends a settlement amount in a dispute, parties might struggle to understand the basis for the recommendation and question its validity.

AI might inhibit our ability to progress our thinking: Reliance on AI discourages critical thinking and innovation in legal practice. For example, lawyers might become overly dependent on AI-generated precedents, stifling creativity and the development of novel legal arguments.

Less or no opportunity to contest decision-making: Opaque AI can lead to a lack of accountability and recourse for individuals affected by AI-generated decisions. For example, a client might find it difficult to challenge an AI-based decision on their case, leaving them without a means to seek redress or alternative outcomes.

Failure to recognize context: Narrow context is a significant limitation of AI systems, which may not fully understand or account for the nuances and complexities of legal cases. For example, an AI tool might overlook critical cultural, social, or historical factors that are essential to a fair and just legal decision.

Conferring rights to a non-living thing: There are ethical and legal questions about the status and responsibilities of AI systems in the legal framework. For example, debates might arise about whether AI systems should have legal standing or accountability in decisions they influence, complicating the legal landscape.

Systems designed without domain expertise: Engineering first can lead to flawed AI applications that do not meet the specific needs and standards of the legal profession. For example, an AI tool developed without input from experienced lawyers might misinterpret legal terminology or processes, leading to inaccurate or unusable outputs.

Lagging regulations to protect traditional rights: There may be gaps in legal protection as AI evolves faster than the regulatory framework, leaving individuals vulnerable. For example, current laws might not adequately address issues of AI accountability or data privacy, leading to legal uncertainties and potential abuses.

Exclusion from benefits, wealth disparity, and economic displacement: AI-driven efficiencies may disproportionately benefit well-resourced entities, exacerbating social and economic inequalities. For example, large law firms might use AI to reduce costs and increase profits, while smaller firms and individual practitioners struggle to compete, widening the economic gap.

Lack of training for younger lawyers: Using and understanding AI tools can create a skills gap, limiting their ability to leverage technology effectively in their practice. For example, law schools might not adequately prepare students for a legal landscape increasingly shaped by AI, putting new graduates at a disadvantage.

Breach of confidentiality: Confidentiality is a significant risk as AI systems handle sensitive legal information, and any breach could have severe repercussions for clients and legal professionals. For example, an AI tool used for case management might be hacked, exposing confidential client information to unauthorized parties and causing significant harm.

Perpetuating existing biases/stagnant law: If AI systems rely on historical data that reflects past biases, they cannot adapt to evolving legal standards and social norms. For example, an AI system might recommend legal strategies based on outdated or biased precedents, hindering progress toward more equitable legal practices.

Dehumanizing law, deemphasizing certain human elements of law: If AI systems prioritize efficiency over empathy, they may ignore the human touch in legal practice. For example, clients might feel alienated or misunderstood if their legal issues are handled primarily by AI systems without meaningful human interaction.

Erosion of trust, accountability, and trust of process: Trust is a significant concern if AI systems are perceived as opaque, unaccountable, or unreliable, undermining confidence in the legal system. For example, if AI-generated legal decisions are not transparent or understandable, clients and the public might lose trust in the fairness and legitimacy of legal outcomes.

Erosion of judgment and expertise: Overreliance on AI diminishes the development and application of human legal judgment and expertise, potentially lowering the overall quality of legal practice. For example, lawyers might become less skilled at critical thinking and complex legal reasoning if they rely too heavily on AI tools for decision-making.

Envisioning Harm Scenarios

From digitization to social media platforms and new advances in AI, the pace and scale of change in the media are unrelenting. The riddle of how these technologies are going to impact us continues to baffle industry leaders and policy makers.

Instead of just speculating about that future, there are adapted tools for systematically exploring the space of the possible. This exercise is exploring with one such tool: **Situation Prototyping**.

In this exercise, we explored how to use **situation prototyping** and creative **scenario-writing** methods to explore the future of the fields, specifically focused on AI technology.

We briefly introduced a small set of technologies, described a scenario-writing method, and put participants into interactive groups to brainstorm and write forward-looking scenarios about the potential impact of the technology on different fields. Each group gathered participants from one of the three fields we were exploring (Medicine, the Law, and Journalism) as well as technologists who are familiar with the technologies and socio-technical metrics for evaluating harm.

Each participant wrote one or more scenarios in their own field. The **scenarios** were short stories that included a setting of time and place, characters with particular motivations and goals, and events that take place. The stories were sequences of character actions and events that logically lead to some **impact of interest** in one of the three fields. Each scenario was focused on one of three technologies:

- **Generative AI**: Generative AI involves creating new content, such as text, images, or music, by leveraging machine learning models trained on vast datasets.
- **Predictive Analytics**: Predictive analytics uses statistical techniques and machine learning algorithms to analyze historical data and make forecasts about future events.
- **Recommendation Engines**: Recommendation engines use algorithms to analyze user preferences and behavior to suggest relevant products, services, or content.

After the breakout groups, we shared and critiqued team scenarios. Participants came away having critically thought about several visions of the future of AI's impact on their field, and, more importantly, with the equipment needed to bring the scenario-writing method with them as they invariably continue to explore the future of impact.

The power of this approach became immediately evident. Participants from each area were able to identify issues that might not have been apparent to those outside that area.

- In Medicine, issues having to do with tiers of service caused by moving between systems, diagnostic issues because of bad training data, and underrepresentation within crucial specialties.
- In Journalism, the erosion of skills leading to errors, the erosion of the standards for news as it drifts towards the popular, and the replacement of editorial decision-making with for-profit optimization.

- In the Law, the creation of an AI-supported generation of new lawyers with the same skills, medicated decision-support undercuts individual advocacy, and autocomplete systems in judicial decision-making can undermine the depth of legal analysis.

Our goal was not complete coverage but rather a proof of concept of the utility of scenario generation as a tool for envisioning issues of harm and impact. To that end, we succeeded and are moving towards the next step of building out simple tools to scale the process. Below, are a few examples of the scenarios:

Medicine

Generative AI (Tiered Medical Services)

Five years from now, medical providers use LLMs only for low-risk activities like scheduling appointments. For diagnosis, SLMs (small language models) trained only with vetted data are used.

The Smith family lives in a small suburb of Chicago. They're a big family with the smallest being 5 years old, and the eldest being 68. Managing doctors' appointments and keeping up prescription refills, vaccine appointments, etc. for everyone can be chaotic. But thankfully, there are personalized apps that automatically schedule appointments and refills!

Granny Smith is 65. She's not very tech savvy, doesn't have much medical knowledge, and is forgetful. She benefits from this app for a few reasons: 1) The app schedules everything. 2) The language used in the app is specialized for her and she can understand it. 3) Her visits are virtual and thus accessible. 4) After each visit, the appointment notes are *"translated"* to language she and her daughter understand in case they need to go back and read the notes.

But even with all these benefits and even though it would be convenient for her to continue using this app, she prefers the *old* ways: in-person appointments and face-to-face interaction with doctors. But her insurance charges a premium for that luxury. Additionally, while the software is convenient, she doesn't like using a screen and is not used to it.

Granny wakes up one day with a terrible cough, which her SLM admits is beyond its 'expertise' - so she turns to an online LLM which suggests she has a grave condition. Worried, she submits a request in her app for an in-person consultation. Dr. John Dole reviews her request, raising his eyebrow upon realizing this diagnosis comes from an inexpert LLM. Wary of headaches associated with "LLM clients'' and LLMs themselves, he defaults to suggesting she rest and check back in two weeks. Granny Joe's condition worsens in the meantime.

Automating medical diagnostics could lead to fragmented care delivery, potentially misguiding patients and creating barriers to timely and accurate medical interventions.

Predictive Analytics (Faulty Training Sets)

One day Bill Howard noticed a dark spot on his skin. Being the busy person that he was, he ignored it for a few weeks. Finally, his wife convinced him to go to the local clinic to get it checked out by a professional. He went to the clinic and found a waiting room packed full of patients who were visibly more ill than he was. He ended up waiting for a few hours before being seen by a doctor who looked exhausted.

Bill explained the situation and the doctor was able to get Bill some imaging done in order to diagnose this spot. The results came back a few minutes after the images were taken and came from a specialized machine learning model trained to diagnose types of skin issues. The doctor immediately agreed with the predicted diagnosis of benign mole, told Bill there was nothing to worry about, and sent him home.

What neither of them knew was that this specialized machine learning model was trained exclusively on images coming from a completely different kind of imaging machine and has picked up on spurious correlations resulting from slight differences in the photo sensors. This has caused the model to be highly inaccurate on images produced by the machine at the clinic Bill went to. The model would therefore always produce a diagnosis that was benign. The overworked and exhausted doctor was unable to spend the time necessary to give a more thorough manual evaluation and relied entirely on this classifier.

Bill noticed that the spot continued to grow over the next few months and grew increasingly worried as well. He went to get a second opinion from a dermatologist who immediately recognized the spot for what it was: skin cancer.

Reliance on a machine learning model trained on images from a different imaging machine led to inaccurate diagnoses due to spurious correlations, endangering patient health.

Recommendation Systems (Skewed Professional Advancement)

James is a second-year medical student who is beginning to think about which clinical rotations he should be taking in his next year of school. He is an imposing figure, six-foot three, and the son of two former professional athletes. He hurt himself hiking during his first year of college and had to pivot from his intense involvement in competitive D1 sports, finding instead a new interest in neuroscience and psychology during undergraduate schooling. During medical school, James, like many other of his co-students, uses an online platform to engage with content during his preclinical years. This platform also considers his extracurricular activities, data from his school record, and shared data from other partner technologies including his web-browsing on campus and purchasing habits. This education platform includes a recommendation algorithm designed to keep students engaged with learning by suggesting courses and clinical rotations that they are likely to enjoy. It also has been shown to increase self-reported satisfaction with clinical specialty selection for graduating medical students.

James has been really getting into biomechanics and musculoskeletal biology courses that have been suggested for him and decides to do his elective clinical rotation in orthopedic surgery. This ultimately influences him to pursue a career in orthopedic surgery, despite having had an interest and aptitude for psychiatry prior to medical school. The recommendation algorithm has been trained on mountains of data of prior medical students from the Northeast of the US and has been found to narrow the demographic diversity of medical subspecialties. For instance, James and other men who engage with sports now make up a higher proportion of orthopedic surgeons, and there are fewer men in the field of psychiatry.

Use of recommendation algorithms in educational platforms can inadvertently narrow the demographic diversity of medical subspecialties by reinforcing existing biases in data.

Journalism

Generative AI (Generation of False News)

Year 2029: Global daily has laid off most of the writers and only has a team of 30 people that take care of writing, marketing, advertisements and fact checking. They have been using GPTUltron for five years and have been able to use it to generate articles to build their customer base and trust by actively doing fact checking. but as the technology was improving exponentially, fact checking has now become just a checkbox to be checked and not something taken seriously as 95% of the time the facts were usually right.

Ana needs to complete an article, but she has been going through a hard time due to her relationship and with no work life balance as now there is no journalist that specializes in one thing but must deal with different sectors. She's burnt out and thinks the LLM will be generating the article anyway so what worse could happen.

Next morning, a major scandal breaks. An AI-generated article falsely accuses a prominent politician of corruption. The news spreads like wildfire, fueled by the AI's propensity to sensationalize. By the time the error is discovered, and a correction is issued, the damage is done. The politician's reputation is tarnished, and public trust in the media hits a new low.

The management at Global Daily gathers for an emergency meeting. The atmosphere is tense as they grapple with the consequences of relying too heavily on AI. As the discussion unfolds, no clear solution emerges. Should they scale back AI usage, risking reduced output and higher costs? Or should they double down, investing in better training and oversight for the AI, hoping to mitigate its flaws?

The story ends with a question mark. The newsroom stands at a crossroads, symbolizing the broader dilemma faced by the media industry. The promise of AI is undeniable, but so are its pitfalls. The future of news remains uncertain, left for the reader to contemplate the path forward.

Over-reliance on AI for news generation, with diminished emphasis on rigorous fact-checking, can lead to significant misinformation and damage to public trust.

Predictive Analytics (Summarization of News)

Bob at Business Insider shows his boss Brenda the startling trend his predictive system had uncovered: over the past ten years, the number of people consuming traditional news stories has drastically decreased, and more and more people are consuming news through summaries on social media, as that is quicker and allows them to argue with strangers online in the comment section. Brenda decides to act quickly and tells her managers to start training their journalists to summarize their articles into a few sentences. One of her managers, Bill, realizes that they can cut costs in his department by getting rid of a few journalists and replacing them with a large language model. Bill's department's articles dominate engagement. Bob's new predictive system, which uses comments on single articles as a proxy for interest, indicates that this is the cheapest way to produce popular and engaging material. Brenda pushes her other managers to follow Bill's lead, and Business Insider faces massive layoffs of journalists, but does hire a few Prompt Engineer Interns.

Five years later, each department now consists of a single journalist, augmented by the Prompt Engineer Interns and the large language model. Each article the journalist produces is summarized and published on various social media platforms. Knowledge about current events is lower than ever, but because the summaries are so short, readers are free to fill in the gaps on their own, which has made Business Insider's social media accounts very popular. Business Insider's profits are higher than ever.

Bill is happy, because in his free time, Bill enjoys arguing with strangers online.

Shifting to AI-generated news summaries for social media can reduce journalistic employment and depth of reporting, potentially lowering public understanding of complex issues.

Recommendation Systems (Information Specialization for Profits)

Michael is 71 years old. He used to get his daily news through Fox News and Facebook. That was five years ago, in 2024. Today, his news is more personalized than ever. Or so he thinks.

The promise of AI-based recommendation systems was that Michael would finally be served exactly the news and analysis that interests and entertains him. This promise has a long history, dating back as far as 1990's "Hyperland," the BBC's documentary-from-the-future.

That promise has always broken down. Instead of personalized agents tailored to each of us individually, the '00s and '10s gave us algorithmic media that served the large platforms. Facebook never figured out what you want or need. Instead, it optimized for whatever content kept you clicking away within Facebook's walled garden.

The AI age replaced the platform era. And, again, it held the promise of perfect, personalized content. But, just below the surface, the AI companies instead were optimizing for whatever-maximizes-share-prices.

Michael is a retiree. He's living on a fixed income. Refinements in digital advertising have made it increasingly clear that he just doesn't inhabit an especially valuable audience segment. He doesn't buy new sneakers or take expensive trips.

And the secret of the AI-based recommendation system is that it isn't optimized for Michael. It's optimized for profit.

Retirees are part of a valuable voting block. They are high-propensity voters. And a small group of high net-worth individuals (tech billionaires, though they prefer you not call them that anymore) heavily subsidizes content that tells the stories they most want to share, framed in ways that are most beneficial to their investment portfolios.

Social scientists at CASMI conduct a massive study of AI-based recommender systems. Their findings reveal that Michael's news diet is, in fact, far less personally-tailored than it seems. He is essentially receiving the stories and frames that Sam Altman and Marc Andreessen and Elon Musk are willing to subsidize. The grand promise of news personalization turns out to be a mirage. The system still bends toward money. And practically all the money is concentrated in a single, sunny California zip code.

CASMI publishes a report, revealing that the impacts of news recommender systems are massively overblown.

Michael never sees the report. Barely anyone does.

AI-based news recommendation systems, while promising personalization, ultimately prioritize profit over individual user needs, shaping content around the interests of powerful stakeholders rather than the diverse needs of the audience.

Law

Generative AI (Leveling of Reasoning)

Context: Five years into the future and Mahlet Velazquez is fresh out of law school and the valedictorian of her class. She just landed a job in a large and prestigious law firm. After her first month, she is invited to compete in an internal mock trial where all participants are given the same legal brief to prepare. This is an account of her participation in this event.

As a former exemplar student, Mahlet prepared to shine in front of her colleagues. To make sure she covered all the required information, she devoted all the evening prior to the competition to study using widely available genAI tools.

Although she was nervous, the competition began. Each student was given the same scenario and asked to present their take on the legal strategy they would suggest to their imaginary client. Participant one took 15 minutes and gave their answer. Participant two gave a very similar answer, which was echoed by participant three, four, five, and by Mahlet. Because they used the same tools, each of their answers were extremely similar.

The law firm partners that were judging this competition had a defeated look. For the past couple of years, the same scenario has been playing out repeatedly. Students were consistently losing their capability for independent or creative solution development. Mahlet's partner mentor, Greg, called her into his office. He told her not to worry. The partners had been experiencing a decrease in new graduate capabilities for some time and were considering funding the opening of a law school that banned the use of generative AI in instruction.

Overreliance on generative AI tools in legal education can lead to a homogenization of thinking and a decrease in the ability to develop independent or creative legal solutions.

Predictive Analytics (Lower Quality of Legal Services)

Alice and Bob have been married for 10 years. They live in Chicago, Illinois with two children (ages 6 and 8). They are seeking divorce as of January 2030.

Alice has been the primary caregiver of the children. Alice and Bob each want primary custody. Alice is being treated for depression.

Alice is employed, but Bob is the primary breadwinner of the household – working for an out-of-state employer – and has a retirement plan.

They both own the marital home with a mortgage and have 100k in assets and 50k in credit card debt.

Both Alice and Bob agree that rather than hiring lawyers, they would preserve the estate in question by using DivorceLawyer.ai to answer their questions, draw up an agreement to be submitted to the court, and save money in the process.

They input the above information into the DivorceLawyer.ai tool. The output agreement assigns primary custody to Bob and only provided Alice with minimal maintenance. While it gave Alice an interest in Bob's retirement plan, it did not advise Alice or Bob about the necessary documentation required to be filed with Bob's employer to protect Alice's interest.

Although Alice is unhappy with the agreement, Bob pressures her to agree and suggests that they will spend all their assets if she does not agree. Feeling that she has no choice, Alice and Bob present the agreement produced by DivorceLawyer.ai to the court as an agreed matter and the court entered judgement based on the agreement.

Relying solely on AI for legal matters such as divorce can lead to inequitable agreements due to a lack of comprehensive legal advice and necessary protections for all parties involved.

Recommendation Systems (Digital Twin Replaces)

As a passionate "early adopter," the Honorable Judge Albert Iglesias was thrilled to pioneer the new Judge Copilot technology from Legal Services, Inc. Judge Copilot (Copilot for short) is billed as a legal doppelganger, loaded with the entire corpus of law in Iglesias's jurisdiction, as well as the content of all his past decisions. Copilot functions as a kind of autocomplete system for judges: by ingesting and processing their past decisions, the judge can see suggestions about how to complete the sentence in front of the cursor.

Iglesias sits down to write his opinion in the case before him, a juvenile charged with driving while under the influence. Copilot begins by identifying similarities and analogies to previous judgments that Iglesias has authored. Once he gets to his ruling, he begins to type, "I find the defendant—" and pauses to reflect.

But a tenth of a second later, a ghostly apparition of text appears in front of the judge's blinking cursor. "...guilty…" Iglesias raises an eyebrow — how did the computer know what he was going to write? And why was it so confident? While Iglesias might agree with the determination, he still finds it off-putting that the machine concluded in a flash without any hint of agonizing or analysis. Nor can Iglesias ask questions of the recommendations or interrogate the machine's confidence levels. The time spent reflecting — in exercising judicial discretion — is collapsed down to a tenth of a second: accept the suggestion or not?

As Iglesias keeps writing, he gets to his sentence. "The defendant is sentenced to—" Copilot suggests, "30 hours of community service." That sounds about right to Iglesias. He hits Tab to accept the suggestion. Copilot continues, "In addition, the defendant is required to complete a DUI education course…" Sure. Iglesias thinks, that sounds like the kind of thing I would say. Tab.

Several Tabs later, and the document has been mostly written by Copilot. Iglesias has blazed through finishing his decision in record time. He looks up from his desk at the library of legal reference books lining his shelves and wonders about the last time he consulted them, poring over them to extract the wisdom contained within. Instead, he has a system that can regenerate his past reasoning in similar cases at lightning speed. But is this an improvement?

The use of AI autocomplete systems in judicial decision-making can undermine the depth of legal analysis and reduce the reflective discretion traditionally exercised by judges, potentially impacting the quality and individual consideration of each case.

Communication Requirements and Plans

To systematically explore how AI impacts are tracked and understood, we conducted an exercise focusing on key questions about information gathering and reporting in various fields. The primary goals were to understand how different fields currently track issues, specifically related to AI, and to identify potential mechanisms for improving AI issue tracking. Participants from various fields, including medicine, law, and journalism, engaged in collaborative discussions to share insights and experiences. They began by examining existing practices in their respective fields, identifying both passive and active forms of information gathering and reporting. This initial step provided a foundation for understanding the current landscape of issue tracking.

The next phase of the exercise involved a deeper dive into AI-specific challenges. Participants explored how their fields are currently addressing AI issues and discussed potential improvements. They considered various mechanisms that might make sense for AI, such as new reporting systems, enhanced data collection methods, and innovative tracking tools. The discussions also included brainstorming on the possible forms this information might take, considering both traditional and digital formats.

We had participants focus on a set of questions related to gathering, organizing, and communicating issues related to harms.

- How do you know what you know?
- What form did your information gathering take? Passive or Active?
- Are there mechanisms for gathering/reporting/collecting?
- How does your field track issues/problems today?
- How is your field tracking issues in AI?
- Are there mechanisms that might make sense for AI?
- What form might the information take?

By the end of the exercise, participants had a comprehensive understanding of existing practices and generated new ideas for effectively monitoring AI-related problems, paving the way for more informed and proactive approaches in their fields.

Medicine

Existing Methods for Tracking AI Harms

In our workshop, medical practitioners and researchers shared various methods they use to stay informed about AI and its associated harms. These methods encompass a range of resources that reflect the intricate and multifaceted nature of healthcare AI. Key sources include the FDA website for post-marketing surveillance, pharmacovigilance pipelines, and data from medical malpractice lawsuits, which provide critical insights into AI-related safety and efficacy. Hospital-based Quality Improvement (QI) tracking systems and the NETS (Nursing Event Tracking System) report help in real-time monitoring and reporting of AI performance in clinical settings. Additionally, the Vaccine Adverse Event Reporting System (VAERS), patient online reports, and outbreak tracking tools offer valuable data on AI impacts. Clinical trials and Phase IV monitoring ensure rigorous assessment of AI technologies, while anonymous reporting systems and SERS (Safety Event Reporting System) facilitate transparency and improvement. Internal corporate data and personal utilization experiences further contribute to identifying and mitigating AI-related harms. This comprehensive collection of resources demonstrates the proactive and thorough strategies employed by healthcare professionals to navigate and address the evolving landscape of AI in medicine.

FDA website post-marketing: The FDA website provides post-marketing surveillance data to track the safety and efficacy of AI systems in healthcare. This resource is essential for doctors and medical researchers to identify potential AI harms after these systems are deployed. For example, updates on AI-related adverse events in medical devices can be found on the FDA's post-market surveillance reports.

Pharmacovigilance pipeline: The pharmacovigilance pipeline is a critical tool for monitoring drug safety and AI-driven drug management systems. It helps in identifying, assessing, and preventing adverse effects or any other drug-related problems. For instance, AI algorithms used in drug prescriptions are monitored for potential errors or biases that could lead to patient harm.

Medical malpractice lawsuits: Information from medical malpractice lawsuits can reveal patterns of AI-related harms in clinical settings. These lawsuits provide case studies and legal precedents that inform safer AI practices. For example, a lawsuit involving an AI diagnostic tool that misdiagnosed a condition highlights the need for better validation and oversight of AI systems.

Hospital based QI tracking: Quality Improvement (QI) tracking systems in hospitals are used to monitor the performance and safety of AI applications in patient care. These systems help identify and rectify issues in real-time, ensuring continuous improvement. For instance, tracking the accuracy of an AI-powered radiology tool to ensure it meets clinical standards.

NETS report (staff reporting system): The NETS (Nursing Event Tracking System) report allows healthcare staff to document and report any AI-related incidents. This reporting system is crucial for early detection and resolution of potential harms caused by AI tools in clinical environments. For example, a nurse reporting an anomaly detected by an AI monitoring system in patient vitals.

Vaccine Adverse Event Reporting System (VAERS): VAERS is a national system for monitoring the safety of vaccines, including those utilizing AI in their development or distribution. This system collects and analyzes data on adverse events following vaccination. For instance, VAERS can be used to track any unexpected reactions linked to AI-driven vaccine deployment strategies.

Patient online reports: Patients often share their experiences and adverse effects of AI-related healthcare interventions through online platforms. These reports provide valuable real-world data for identifying potential harms and improving AI systems. For example, patients reporting side effects from an AI-recommended medication regimen.

Outbreak tracking: AI tools used in outbreak tracking help monitor and predict the spread of infectious diseases. These systems can identify potential AI-related errors in data collection and analysis that could impact public health responses. For instance, an AI model incorrectly predicting the spread of a disease due to biased data inputs.

Clinical trials: Clinical trials involving AI technologies are rigorously monitored to assess their safety and efficacy. These trials provide critical data on the potential harms and benefits of AI applications in medicine. For example, a clinical trial evaluating an AI system for early cancer detection.

Anonymous reporting for bias: Anonymous reporting systems allow healthcare professionals to report AI biases without fear of retaliation. This promotes transparency and helps address ethical concerns related to AI use in medicine. For example, a doctor anonymously reporting bias in an AI algorithm used for patient triage.

SERS electronic reporting and hospital event: The Safety Event Reporting System (SERS) enables electronic reporting of AI-related safety events within hospitals. This system facilitates timely intervention and improvement of AI tools. For instance, reporting a malfunction in an AI-driven surgical robot.

Phase IV monitoring: Phase IV monitoring involves post-marketing surveillance of AI technologies used in healthcare to ensure long-term safety and effectiveness. This phase is crucial for detecting any delayed adverse effects or issues that arise after widespread use. For example, ongoing monitoring of an AI-based diagnostic tool for potential long-term impacts.

Internal/private corporate data: Companies often collect and analyze internal data on the performance and safety of their AI products. This data helps in identifying and mitigating potential harms before they reach the market. For example, a tech company conducting internal audits of their AI healthcare applications.

Personal utilization: Personal experiences and utilizations of AI tools by healthcare professionals provide practical insights into their efficacy and potential harms. These first-hand accounts are invaluable for continuous improvement and user feedback. For instance, a physician sharing their experience with an AI-powered clinical decision support system.

Information Requirements/Suggestions

The following list presents the methods suggested by medical researchers for obtaining information about AI-related harms. These recommendations were developed through a collaborative process of individual suggestions followed by group curation. A centralized clearinghouse consolidating consensus activities would serve as a comprehensive resource, streamlining access to best practices, regulatory updates, and expert opinions. Incentives for data sharing and hypothesis testing can enhance collaboration and innovation, encouraging researchers to contribute to shared databases. Establishing a common vernacular for AI in healthcare can improve communication and collaboration. Ensuring AI-related information is easily searchable enhances accessibility for medical researchers. Requiring AI models to include 'limitations' or 'issues' documentation promotes transparency and informed usage. Active documentation of AI-related incidents by research communities provides valuable insights into real-world challenges. Centralized standards administered by federal or state governments ensure consistent regulations across the healthcare system. Anonymizing data allows for comprehensive research while protecting patient privacy. Legal protection of shared data encourages broader data sharing and collaboration. Communicating AI-related information in an understandable way to the public improves trust and engagement. Developing a taxonomy of AI use cases helps guide monitoring and evaluation efforts. Financial incentives motivate researchers to engage in AI studies and share their findings. Access to curated sources of information about AI harms is essential for providing well-rounded and critical perspectives on AI technologies.

Clearinghouse/consensus activities: A centralized clearinghouse consolidating consensus activities related to AI in healthcare would provide a comprehensive resource for medical researchers. This platform would streamline access to best practices, regulatory updates, and expert opinions. For example, a researcher might access a clearinghouse to find consensus guidelines on AI implementation in diagnostic imaging.

Incentives for data sharing and/or hypothesis testing: Offering incentives for data sharing and hypothesis testing can boost collaboration and innovation among medical researchers. These incentives could include grants, recognition, or access to shared databases. For instance, a researcher might receive funding for sharing data on AI-driven clinical trials, facilitating broader hypothesis testing.

Common vernacular/shared landscape: Developing a common vernacular and shared landscape for AI in healthcare can enhance communication and collaboration among researchers. This standardized language would help ensure clarity and consistency across studies and publications. For example, researchers from different institutions could collaborate more effectively using a shared terminology for AI-related concepts.

Make easily searchable: Ensuring that AI-related information is easily searchable can significantly improve accessibility for medical researchers. This involves creating user-friendly databases and search tools that allow quick retrieval of relevant data. For instance, a researcher might use an advanced search engine to locate specific studies on AI applications in oncology.

AI models need “limitations” or “issues” documentation: Requiring every AI model to include an easily understandable "limitations" or "issues" document can enhance transparency and informed usage. This documentation would help researchers and practitioners understand the boundaries and potential pitfalls of AI tools. For example, a researcher might review the limitations document of an AI diagnostic tool to assess its suitability for a specific patient population.

Research communities and forums actively documenting incidents: Active documentation of AI-related incidents by research communities and forums can provide valuable insights into real-world challenges and solutions. This collective knowledge can inform best practices and safety protocols. For instance, a researcher might consult a forum to learn about incidents of bias in AI algorithms and strategies for mitigation.

Centralized standards (i.e. federal or state government): Centralized administration of AI programs by federal or state governments can ensure consistent standards and regulations across the healthcare system. This oversight can enhance accountability and coordination. For example, a researcher might refer to government-administered guidelines for ethical AI use in clinical trials.

Anonymization of data so information can come from public and private: Anonymizing data allows for the inclusion of information from both public and private sources, fostering comprehensive research. This practice protects patient privacy while enabling robust data analysis. For instance, a researcher might analyze anonymized patient data from multiple hospitals to study AI's impact on treatment outcomes.

Legal protection of shared data: Providing legal protection for shared data can encourage more researchers to contribute their findings without fear of misuse or litigation. This legal framework can facilitate broader data sharing and collaboration. For example, a researcher might share clinical trial data on an AI platform knowing that it is legally protected.

Find way to share information in understandable way for public: Sharing AI-related information in a way that is understandable to the public can bridge the gap between researchers and the general population. This approach can improve public trust and engagement in AI advancements. For instance, a researcher might collaborate with science communicators to create accessible summaries of their AI research findings.

Taxonomy of use cases to help guide monitoring: Developing a taxonomy of AI use cases can help guide monitoring and evaluation efforts. This structured classification can provide clear criteria for assessing AI applications in healthcare. For example, a researcher might use a taxonomy to categorize and monitor different AI tools used in patient diagnostics.

Financial incentives: Financial incentives can motivate researchers to engage in AI studies and share their findings. These incentives might include grants, awards, or funding for collaborative projects. For instance, a researcher might receive a grant to develop and test a new AI model for predicting disease outbreaks.

Access to curated sources of harm associated with AI: Access to curated sources of information about harms associated with AI is essential for researchers. These sources compile data and case studies on the negative impacts of AI, such as biases, job displacement, and privacy violations. By referencing these curated sources, researchers can provide well-rounded and critical perspectives on AI technologies. For example, a researcher might use a curated database to highlight instances of AI-induced job losses in the healthcare sector.

Centralized database that collects all the existing information services: A centralized database that collects all the existing information services is essential for ensuring that researchers have access to comprehensive and up-to-date information. This database would help in identifying patterns and trends related to AI harm and enable more informed decision-making. For example, if multiple hospitals report similar adverse events related to a particular AI algorithm, this database will help identify and address the issue promptly.

Phase IV style registry of deployed SaMD: A Phase IV style registry of deployed Software as a Medical Device (SaMD) would track the performance and safety of these technologies in real-world settings. This registry would provide valuable post-market surveillance data to detect any emerging issues or adverse effects. For instance, a registry could reveal that a new diagnostic AI tool is less accurate in certain demographic groups, prompting further investigation and improvement.

Mandated reporting for adverse effects: Mandated reporting for adverse effects ensures that any negative outcomes associated with AI in healthcare are promptly and systematically reported. This process would help in the timely identification and mitigation of potential harms, improving overall patient safety. For example, a mandated report might uncover an AI system that frequently misinterprets medical images, leading to incorrect diagnoses.

Centralized analysis of possible harms: Centralized analysis of possible harms involves a dedicated team or entity analyzing data from various sources to identify potential risks associated with AI. This centralized approach allows for a more coordinated and comprehensive understanding of the impact of AI technologies. For example, centralized analysis might identify that a particular AI-driven treatment plan increases the risk of certain side effects in patients with pre-existing conditions.

Public-private data sharing: Public-private data sharing facilitates the exchange of information between government agencies, private companies, and research institutions. This collaboration can enhance the quality and breadth of data available for assessing AI-related risks and benefits. For instance, data sharing between a private AI company and a public health agency could lead to the early detection of an AI system's bias against minority groups.

AI harm/AI help web site: An AI harm/AI help website would serve as a resource for healthcare professionals and the public to report and learn about AI-related issues. This platform could provide educational materials, support resources, and a forum for sharing experiences and solutions. For example, a clinician could use the website to report a malfunctioning AI tool and find guidance on alternative solutions.

Mandated reporting by insurance companies, hospitals, etc.: Mandated reporting by insurance companies, hospitals, and other healthcare entities ensures that a wide range of stakeholders contribute to the monitoring of AI impacts. This comprehensive reporting can lead to a more accurate and complete picture of AI-related harms and benefits in the healthcare system. For example, an insurance company might report an increase in claims related to an AI-driven surgical robot, prompting a review of its safety.

Journalism

Existing Methods for Tracking AI Harms

In our workshop, journalists shared various methods they use to stay informed about AI and its associated harms. These methods encompass a wide range of sources and interactions, highlighting the multifaceted approach journalists take to cover this complex field. From informal word-of-mouth exchanges and direct conversations with experts to specialized beats and reader feedback, journalists gather insights from diverse perspectives. They also leverage social media platforms, traditional news sources, press releases, and academic publications to stay current with AI developments. Additionally, attending conferences and using AI technologies themselves provide firsthand experiences that enrich their reporting. This compilation of resources demonstrates the dynamic and comprehensive strategies journalists employ to navigate and report on the evolving landscape of AI and its impacts.

Word of Mouth: Journalists often rely on word of mouth to stay informed about AI issues. This involves hearing about developments and concerns through informal conversations and personal networks. For example, a journalist might learn about a new AI startup from a colleague during a casual lunch meeting.

Interviewing People: Direct conversations with experts, stakeholders, and the general public provide journalists with diverse perspectives on AI. These interactions help them understand the broader implications of AI technologies. For instance, interviewing a data scientist can offer insights into the technical challenges of implementing AI.

This is a journalist's beat: Some journalists specialize in covering AI as their dedicated beat. This specialization allows them to develop deep expertise and stay up-to-date with the latest advancements and controversies in the field. For example, a tech journalist might consistently report on AI ethics and regulation.

Reader Feedback: Feedback from readers plays a crucial role in informing journalists about AI-related issues. Comments, emails, and social media interactions can highlight public concerns and interests. For instance, a surge of reader comments on an AI article might prompt further investigation into the topic.

Social media: Social media platforms are a valuable source of information for journalists covering AI. These platforms provide immediate access to breaking news, expert opinions, and community reactions. Journalists use these platforms to track trending topics, follow influential voices, and gauge public sentiment. For example, a journalist might discover a viral AI video on TikTok that sparks a new story idea.

Social Media Listening: Journalists use social media listening tools to analyze online conversations about AI. This helps them identify emerging trends, popular topics, and key influencers in the field. For example, a journalist might use a tool to track mentions of AI bias in social media posts.

News/magazine sources/articles: Traditional news sources and specialized magazines provide journalists with in-depth analysis and reports on AI. These publications offer well-researched articles that enhance journalists' understanding of complex issues. For example, a journalist might cite a detailed AI report from a tech magazine in their article.

Established Sources: Established sources, such as industry insiders and experts, provide journalists with reliable information about AI. These sources often offer insights that are not publicly available. For example, an industry insider might reveal details about a confidential AI project to a journalist.

News sources: Journalists monitor news outlets for reports on layoffs and other significant events in the AI industry. These stories can signal broader trends and shifts within the sector. For example, news of layoffs at a major AI company might indicate financial troubles or strategic changes.

Public editors/tech reporters: Public editors and technology reporters serve as valuable resources for journalists covering AI. They often have specialized knowledge and can provide critical analysis and commentary on AI developments. For example, a tech reporter might explain the implications of a new AI regulation.

Press releases/policy organizations: Press releases from companies and policy organizations inform journalists about new AI products, research, and regulatory changes. These official statements help journalists stay current with industry announcements. For instance, a journalist might write about a new AI tool after receiving a press release from the developer.

Fact-checking organizations: Fact-checking organizations play a vital role in verifying information related to AI. Journalists rely on these organizations to ensure the accuracy of their reporting. For example, a journalist might use a fact-checking service to verify claims about AI's impact on employment.

Journalists who observe & share incidents informally: Some journalists observe AI-related incidents informally and share their findings through blogs, social media, or personal networks. These observations can reveal important, often overlooked, aspects of AI deployment. For example, a journalist might blog about a personal experience with an AI-powered customer service bot.

Conferences: Attending AI conferences allows journalists to network with experts, attend presentations, and gather firsthand information. These events provide insights into the latest research, technologies, and industry trends. For instance, a journalist might report on new AI innovations showcased at a major tech conference.

Through use: Journalists also gather information about AI by using the technologies themselves. This hands-on experience allows them to understand the practical applications and limitations of AI tools. For example, a journalist might test an AI-driven photo editing software to evaluate its capabilities.

News rating systems (e.g., NewsGuard): News rating systems like NewsGuard help journalists assess the credibility of sources reporting on AI. These ratings guide journalists in distinguishing reliable information from misinformation. For instance, a journalist might use NewsGuard to verify the trustworthiness of a news website before citing it in their article.

Academics/scholarly papers: Academic publications and scholarly papers offer in-depth research and theoretical insights into AI. Journalists use these resources to understand the scientific and ethical dimensions of AI technologies. For example, a journalist might reference a scholarly paper on AI ethics in their reporting.

Information Requirements/Suggestions

The following list outlines the methods suggested by journalists for obtaining information about AI-related harms. These recommendations were formulated through a process of individual input and group curation. Journalists emphasize the importance of industry group issues and updated best practices to stay informed about the latest technological and ethical standards. They advocate for mandated transparency requirements for tech companies, ensuring detailed disclosures about AI systems and their biases. Enhanced training on AI fundamentals is essential for journalists to improve their understanding and reporting on complex AI topics. Studies and polls assessing public trust in facts and media are valuable for gauging public sentiment and building audience trust. Independent professional certification authorities for AI reporting can enhance journalistic credibility. Increased collaboration with academia provides access to cutting-edge research and expert analysis, ensuring well-informed coverage. Engaging with grassroots movements offers unique perspectives on the impact of AI on marginalized communities. Specialized training courses on AI, including technical, ethical, and regulatory aspects, are crucial for improving reporting skills. Access to curated sources of information about AI harms, such as biases, job displacement, and privacy violations, enables journalists to provide critical and comprehensive perspectives on AI technologies.

Industry group issues/updated best practices: Journalists often seek information from industry groups that issue guidelines and best practices for AI. These groups provide updated recommendations that reflect the latest technological advancements and ethical considerations. This helps journalists stay informed about industry standards and changes. For instance, a journalist might refer to best practice guidelines from the Partnership on AI when writing about AI ethics.

Mandated transparency requirements for tech companies: Journalists would benefit from mandated transparency requirements imposed on tech companies. These requirements would compel companies to disclose detailed information about their AI systems, including how they function and their potential biases. This transparency enables journalists to report more accurately and critically on AI technologies. For example, a journalist might use disclosed data from a tech company to investigate biases in an AI hiring tool.

More training on AI fundamentals: To improve their reporting, journalists need more training on AI fundamentals. This training would cover the basic concepts, technologies, and ethical considerations of AI. With a solid understanding of AI, journalists can better analyze and explain complex AI topics to their audience. For example, a journalist might take an online course on machine learning to better understand the subject and improve their reporting.

Studies/polling of state of public trust in facts/media: Studies and polls that assess public trust in facts and media are valuable resources for journalists. These studies provide insights into how the public perceives AI and media coverage of AI. Understanding public sentiment helps journalists address concerns and build trust with their audience. For instance, a journalist might cite a poll showing declining trust in media when discussing the importance of transparency in AI reporting.

Independent professional certification authorities: The establishment of independent professional certification authorities for AI could enhance journalists' credibility. These authorities would certify journalists who demonstrate a thorough understanding of AI and ethical reporting practices. Such certification could signal to readers that a journalist is knowledgeable and trustworthy. For example, a journalist might pursue certification from a recognized AI ethics organization to bolster their credentials.

More collaboration with academia (e.g., The Markup): Journalists would benefit from increased collaboration with academia on AI-related topics. Partnerships with academic institutions can provide journalists with access to cutting-edge research and expert analysis. This collaboration helps ensure that media coverage of AI is well-informed and grounded in scientific evidence. For example, a journalist might work with researchers from The Markup to investigate algorithmic bias in social media platforms.

Grassroots: Grassroots movements can offer journalists unique perspectives on AI issues. These movements often highlight the impact of AI on marginalized communities and advocate for ethical practices. By engaging with grassroots organizations, journalists can amplify underrepresented voices and bring attention to important social issues. For instance, a journalist might report on a grassroots campaign advocating for greater AI accountability in policing technologies.

Training courses: Journalists need access to specialized training courses on AI to enhance their reporting skills. These courses could cover a range of topics, from technical aspects of AI to ethical considerations and regulatory frameworks. With comprehensive training, journalists can produce more accurate and insightful reports on AI. For example, a journalist might enroll in a training course on AI ethics to better understand the moral implications of AI technologies.

Access to curated sources of harm associated with AI: Access to curated sources of information about harms associated with AI is essential for journalists. These sources compile data and case studies on the negative impacts of AI, such as biases, job displacement, and privacy violations. By referencing these curated sources, journalists can provide well-rounded and critical perspectives on AI technologies. For instance, a journalist might use a curated database to highlight instances of AI-induced job losses in the manufacturing sector.

Law

Existing Methods for Tracking AI Harms

The following list provides an insightful overview of the various sources of information that legal scholars and lawyers rely on to stay informed about the potential harms associated with artificial intelligence (AI). These sources were identified through a collaborative process, beginning with individual suggestions and culminating in a curated compilation by the group. Legal professionals often turn to well-funded investigative journalism, such as ProPublica’s reports on predictive policing biases, and weekly newsletters like 'Algorithmic Justice' for regular updates. Traditional news segments, whistleblowers, and post-accident reports from organizations like the NTSB offer crucial insights into specific incidents of AI failure. Additionally, academic research papers, conferences, and professional guidelines contribute scholarly perspectives and best practices. Legal publications, internal reports, and practical advice from legal clinics provide detailed analyses and support. The daily WIRED newsletter and discussions within the Data Science Community Network keep legal professionals abreast of the latest technological developments and challenges. Social media platforms, personal networks, and communities like Law X (formerly Twitter) facilitate real-time information exchange. Mainstream news outlets, legal ethics roundups, and word-of-mouth communications also play significant roles in disseminating knowledge about AI harms. Finally, personal experiences with AI systems further enrich the understanding of their potential impacts.

Well-funded investigative journalism: Legal scholars and lawyers often rely on well-funded investigative journalism to uncover detailed information about AI harms. For example, ProPublica's investigation into the biases of predictive policing algorithms has provided critical insights into how these systems can disproportionately affect minority communities.

Once-per-week email newsletter: A once-per-week email newsletter can be a valuable resource for staying updated on the latest developments regarding AI harms. For instance, the 'Algorithmic Justice' newsletter summarizes the most important news and studies related to AI ethics and safety, helping professionals stay informed without needing to sift through multiple sources.

Segment of traditional news: Traditional news segments often highlight significant incidents of AI harm, bringing them to the attention of a broader audience. For example, a CNN segment on a self-driving car accident raised awareness about the potential dangers and regulatory gaps in autonomous vehicle technology.

Whistleblower: Whistleblowers play a crucial role in revealing hidden AI harms that might otherwise go unnoticed. A notable example is the case of Timnit Gebru, a former Google researcher who exposed the ethical issues within Google's AI research division, leading to broader discussions about AI ethics in the industry.

Post AI harms/accident reports (NTSB): Reports from organizations like the NTSB following AI-related accidents offer detailed insights into what went wrong. For example, the NTSB's report on the fatal Uber self-driving car crash provided an in-depth analysis of the technological and human failures involved.

Research papers: Academic research papers provide a scholarly perspective on AI harms, often backed by rigorous methodologies and extensive data. For instance, a research paper on the adversarial vulnerabilities of facial recognition systems can help lawyers understand the technical flaws and potential legal implications.

Academic convenings: Conferences and symposiums bring together experts to discuss and debate the latest findings on AI harms. An example is the annual AI Now Symposium, where researchers, policymakers, and legal professionals gather to share their latest findings and insights on AI's societal impacts.

Professional guidance: Professional bodies and associations often release guidelines and best practices for dealing with AI harms. For example, the American Bar Association's guidelines on AI and ethics help lawyers navigate the complexities of AI-related cases with informed strategies.

Legal publications: Journals and legal magazines publish articles and case studies on AI harms, providing detailed legal analysis. For example, the Harvard Law Review might feature a comprehensive article on the legal challenges of regulating autonomous weapons systems.

Internal reports: Companies and organizations may produce internal reports detailing instances of AI harm and their responses. For example, an internal audit report from a tech company could reveal how an AI system's bias led to discriminatory hiring practices, providing key evidence for legal professionals.

Legal clinics & help desk services: Legal clinics and help desks provide practical advice and support on AI-related issues, often dealing directly with affected individuals. For instance, a legal clinic specializing in technology law might assist a client whose privacy was violated by a flawed AI system.

WIRED daily newsletter: The WIRED daily newsletter offers timely updates on technology, including AI and its associated harms. An example might be a daily update featuring a story about an AI glitch that led to widespread data breaches, highlighting the need for better security measures.

DSCN Data science community network: The DSCN Data Science Community Network facilitates discussions and knowledge sharing about AI harms among data scientists and technologists. For instance, a discussion thread about an AI model's failure in healthcare settings can provide valuable insights for legal scholars studying medical AI regulations.

Social media such as X/Reddit/LinkedIn: Social media platforms like X (formerly Twitter), Reddit, and LinkedIn are vibrant spaces for real-time discussions about AI harms. For example, a trending thread on Reddit might discuss the ethical implications of a new facial recognition technology, providing anecdotal evidence and public sentiment.

Lawyer friends: Personal connections with other lawyers can be an invaluable source of information about AI harms. For instance, a lawyer might learn about a pending lawsuit involving AI bias in loan approvals through a casual conversation with a colleague who is working on the case.

Colleagues connected to technology companies: Colleagues working in technology companies often have first-hand knowledge of AI systems and their potential harms. For example, a friend working at a tech startup might share insider information about the unintended consequences of their new AI product, providing valuable context for legal analysis.

Law X: Law X (formerly Twitter) is a niche community where legal professionals discuss and debate issues, including AI harms. An example is a viral tweet thread discussing the recent legal challenges faced by an AI company accused of privacy violations, which helps lawyers stay updated on current legal discourse.

CNN/NYT 'News': Mainstream news outlets like CNN and The New York Times frequently cover significant AI harm incidents. For example, an investigative article in The New York Times about the flaws in AI-driven hiring processes can shed light on systemic issues and spur legal debates.

Legal ethics roundup (Renee Knake Jefferson): Legal ethics roundups, such as those curated by Renee Knake Jefferson, provide critical insights into ethical considerations surrounding AI harms. For example, a roundup might include a case study on the ethical dilemmas faced by lawyers representing clients affected by biased AI systems.

Word of mouth: Informal conversations and word of mouth are often the quickest ways to learn about new instances of AI harm. For instance, a lawyer might hear about a recent AI-related incident at a networking event, prompting them to investigate further.

Personal experience: Direct personal experience with AI systems and their failures provides a unique and firsthand perspective on AI harms. For example, a lawyer who has worked on a case involving a malfunctioning AI medical device can draw on their experience to better understand and address similar issues in the future.

Information Requirements/Suggestions

The following list represents the improved strategies suggested by legal scholars for obtaining information about AI harms. These recommendations were developed through a collaborative process that involved individual suggestions followed by group curation. Legal professionals advocate for robust breach and incident reporting laws to enhance transparency and accountability. The importance of competent investigators to uncover AI-related harms is emphasized, along with the need for incentives to promote good AI governance. Professional organizations are highlighted for their role in monitoring AI use in law, providing guidelines, and offering training. Funding is seen as crucial for supporting social scientists and journalists to study and report on AI's societal impacts. An American Bar Association (ABA)-managed repository of AI incidents is proposed to centralize data on AI-related harms. Standards bodies and professional organizations are encouraged to establish repositories and guidelines while incentivizing best practices. Sociotechnical audits are recommended to assess both the technical and social impacts of AI systems. Legal malpractice insurers, regular columns in legal media, and destigmatizing the reporting of negative AI results are suggested to promote responsible AI use and transparency. Protections for whistleblowers and procurement processes that incorporate ongoing oversight are essential for accountability. Morbidity and Mortality (M&M) conferences adapted to discuss AI-related harms, in-depth legal analysis of AI implications, and the establishment of an authoritative body to collect and analyze AI-related harms are also recommended to ensure comprehensive oversight and informed policymaking.

Breach/incident reporting laws: Lawyers want robust breach and incident reporting laws to ensure transparency and accountability when AI systems fail. These laws would mandate the disclosure of AI-related incidents, helping legal professionals track and respond to harms more effectively. For example, a law requiring companies to report any AI system malfunctions that result in consumer harm.

Investigators: Competent investigators are crucial for uncovering the details of AI-related harms. Legal professionals rely on investigators to gather evidence and provide expert analysis on AI incidents. For example, hiring a specialist to investigate a case where an AI algorithm led to wrongful termination of employees.

Incentives/carrots and sticks: Incentivizing good practices and penalizing negligence can drive better AI governance. This approach encourages organizations to adopt safer AI practices while holding them accountable for failures. For instance, offering tax breaks to companies that implement comprehensive AI safety protocols.

Competent & Attentive Professional Organizations monitoring AI in law: Professional organizations play a key role in monitoring AI use in the legal field. They provide guidelines, training, and oversight to ensure ethical and competent use of AI. For example, the ABA setting up a dedicated committee to monitor AI applications in legal practice.

Funding Social scientists to study AI: Funding social scientists to research AI's societal impacts can provide valuable insights for legal professionals. For instance, grants supporting studies on AI bias in judicial decisions.

Funding Journalism: Supporting journalism to investigate AI harms ensures that these issues receive public attention. For example, funding investigative pieces on AI misuse in surveillance technologies.

Funding for investigative journalism: Specific funding for investigative journalism can uncover hidden AI-related issues. For instance, grants for journalists to explore the impacts of AI in healthcare.

ABA repository of incidents: An ABA-managed repository of AI incidents would serve as a centralized database for tracking and analyzing AI-related harms. Legal professionals could use this resource to study past cases and develop better strategies for future incidents. For example, a searchable database detailing various AI-induced legal cases and their outcomes.

Standards bodies and professional organizations (repositories, guidelines, incentivization): Standards bodies and professional organizations can establish repositories and guidelines while incentivizing best practices. These entities provide a framework for responsible AI use in the legal field. For instance, the Institute of Electrical and Electronics Engineers (IEEE) creating standards for ethical AI use in law.

Sociotechnical audits: Conducting sociotechnical audits helps assess both the technical and social impacts of AI systems. These audits provide a comprehensive understanding of how AI affects various stakeholders. For example, auditing an AI-driven hiring platform to evaluate its impact on diversity and inclusion.

Legal malpractice insurers: Legal malpractice insurers can play a role in promoting responsible AI use by setting insurance requirements. This could include mandating certain safeguards and practices for AI-related services. For instance, insurers requiring law firms to implement bias mitigation strategies in AI tools.

A regular column in prominent legal media: A regular column in legal media can keep the community informed about AI harms and developments. This platform allows for ongoing discussion and dissemination of important findings. For example, a monthly column in the ABA Journal discussing recent AI-related legal cases.

Destigmatize negative results: Destigmatizing the reporting of negative AI results encourages transparency and learning from failures. This approach helps build a culture of openness and continuous improvement. For instance, promoting case studies that highlight both successes and failures of AI implementations in legal contexts.

Protections for whistleblowing: Legal protections for whistleblowers are essential for uncovering AI harms. These protections ensure that individuals can report issues without fear of retaliation. For example, laws that safeguard employees who expose unethical AI practices within their organizations.

Procurement processes that incorporate ongoing oversight & disincentives overpromising: Procurement processes should include mechanisms for ongoing oversight and penalties for overpromising AI capabilities. This ensures that AI vendors remain accountable throughout the lifecycle of their products. For instance, requiring regular performance reviews and audits for AI systems used in public services.

M&M conferences: Morbidity and Mortality (M&M) conferences can be adapted to discuss AI-related harms, providing a forum for learning and improvement. These conferences allow professionals to review and analyze AI incidents in a constructive manner. For example, an annual M&M conference focusing on AI errors in legal practice and their resolutions.

Legal analysis of implications of AI harms: In-depth legal analysis of AI harms helps professionals understand the broader implications of these technologies. This analysis can inform policy-making and legal strategies. For instance, publishing a comprehensive report on the legal ramifications of AI-induced discrimination in employment.

Use of an authoritative authority collecting harms: Establishing an authoritative body dedicated to collecting and analyzing AI-related harms would provide a central point of reference for legal professionals. This authority could compile data, issue reports, and recommend best practices based on comprehensive research. For example, a government agency that collects and publishes detailed reports on AI-induced harms across various industries.

Outcomes

The workshop explored AI's impact in Medicine, Law, and Journalism by bringing together researchers and practitioners to identify risks and develop targeted strategies to address them. Our goal was to facilitate domain-specific exploration of harms and benefits, gathering the information needed to address AI's challenges with tailored, actionable plans for managing its effects.

Process: The process was designed around the idea of using our domain experts as an information resource. The workshop was structured around three main components:

1. **Mixed-Field Panel:** The workshop began with a panel featuring professionals from Medicine, Law, Journalism, and technology sectors. This panel outlined the major challenges and themes related to AI, setting a foundation for the subsequent discussions.
2. **Framing Talks:** Before each breakout session, experts delivered talks to establish a common understanding of the key issues within each of the breakout sessions.
3. **Breakout Sessions:** Participants were organized into groups to focus on three specific issues:
    - **Identifying Existing Harms:** This breakout session focused on pinpointing specific positive and negative impacts of AI within each domain.
    - **Anticipating Future Harms:** This breakout concentrated on anticipating future issues involving AI in each domain through scenario generation.
    - **Exploring Safety Communication Strategies:** This breakout was aimed at the requirements for effective communication plans related to the risks, benefits, and impact of AI for each domain and the broader public.

Of course, after the breakout sessions, participants reconvened to present their findings. This collaborative discussion aimed to synthesize insights into practical strategies for addressing AI-related risks in each field.

Outcomes: By engaging domain-specific practitioners and exploring real-world impacts, we were able to generate concrete results that address the challenges posed by AI in a practical and targeted way. These outcomes include:

1. **Compilation of Domain-Specific Harms and Benefits:** We developed specific examples of AI's harms and benefits for Medicine, Law, and Journalism which we can now use to inform ethical best-practices and guide responsible AI use in each domain.
2. **Validation of Envisioning Techniques for Problem Anticipation:** Participants created a suite of scenarios that mapped out AI's potential long-term effects and demonstrated the effectiveness of scenario-writing as a tool for anticipating future challenges.

3. **Enhanced Understanding of Communication Methods:** Finally, we identified current communication methods within each domain, highlighting their utility and shortcomings, and developed a set of strategies to better convey AI-related risks and benefits.

Next Steps

The overall goal of CASMI's work is to develop an understanding of the possible harms associated with AI, methods for both mitigating and avoiding them, and a communication strategy that attends to different audiences, their concerns, and their information needs. As part of this process, CASMI has been working on studies derived from existing examples of harms produced by decisions in the design, development, and deployment of AI systems to provide decision-makers with examples that illuminate how certain practices may lead to unwanted and harmful outcomes.

The target for this workshop was requirements gathering needed to shape a plan for the right communication at the right time for the right constituencies.

With these requirements in hand, the next steps comprise three elements: methods for domain-level case gathering and curation, techniques for case validation and analysis at scale, and approaches to domain- and role-centric document generation based on the corpus we are developing.

Case gathering and curation

The workshop made it clear that each of our different domains has a different set of standards for tracking harm. In Medicine, Incident Reporting Systems **and** Patient Safety Indicators are linked in the capture and recording of harm. In the Law, Ethics and Professional Responsibility Complaints and Legal Audits play the same role. In Journalism, Fact Checking and Verification are done both internally and externally. While different, each area has an initial mechanism for identification that can be used as a feeder into the AI Incident Database (AIID) and a source for cases.

In parallel, our envisionment exploration made it clear that we can employ scenario-building as a mechanism for problem anticipation and use the results as part of the AIID. The resulting examples would provide a wider view of what harms could be alongside the harms we have already seen.

In service of this, we propose two steps:

> First, approach the industry organizations for each of these fields with a plan for widespread gathering of examples based on existing protocols.
>
> Second, provide them with a tool that supports scenario building that can be sent to their members for the generation of examples in anticipation of harm.

Validation and Analysis

The goal in this phase is to understand different failures that have occurred and then transform those insights into different communication vehicles for the different stakeholders. Our approach to this is the use of a crowd-sourced approach that draws on editorial and analytical skills linked to each of the fields. This approach opens the door to using the schools associated with each of the fields.

In service of this, we propose development of a domain-centric crowd-sourced approach that is linked to the different schools related to the different fields. Our starting point for this would be Northwestern’s schools of Law, Medicine, and Journalism with the plan of expansion once we establish the mechanisms for review.

Domain- and Role-Centric Document Generation

Our view has always been that the best way to communicate issues of harm are through focused generation of different documents that might communicate the same ideas but do so for different stakeholders, their information needs, and their role in decision-making. This workshop confirmed and amplified this belief. As we move forward, we are looking at developing a cadre of editorial staff (paid, volunteer, student) to craft variants of documents different stakeholders and domains.

In service of this, we propose the development of an editorial workforce aimed at generating:

- Case studies for business stakeholders
- Design documents for developers
- Policy briefs for regulators and thought leaders with different domains
- Domain focused descriptions for practitioners

Our goal in this work is to craft an open development model that scopes across the data gathering process through analysis and dissemination to put the right information in the right form in the hands of the decision makers and developers who need it. In doing this, we see a distributed approach that utilizes multiple techniques at each phase. We envision a responsive model that supports a constant flow of multiple inputs, flexible process models, and multiple outputs that are variants of a single product vision.