Northwestern | CENTER FOR ADVANCING SAFETY OF MACHINE INTELLIGENCE

Roadmap for Research on the Human Impact of Machine Learning Applications

Center for Advancing Safety of Machine Intelligence

Version 1.1

August 2022

This work was supported through funding from Underwriters Laboratories.

Executive Summary

Machine Learning (ML) and Artificial Intelligence (AI) are transformational technologies fueling gains in everything from processing data to fighting fires. Medical diagnostic systems, autonomous vehicles, and personal assistants have all incorporated the power of AI, but its rapid adoption increasingly outpaces our ability to understand its human impact. While many groups are focused on the fundamental articulation of ethical principles for AI, there is an urgent need for mechanisms to operationalize the principles of ethics in system development.

Northwestern University and UL Research Institutes established the **Center for Advancing Safety of Machine Intelligence (CASMI)** in response to this need. **CASMI** is a research hub founded to create best practices for safely designing, developing, and deploying Artificial Intelligence technologies. To further its mission of operationalized safety, CASMI is creating tools for evaluation and is identifying the necessary further research.

This **Research Roadmap** encompasses approximately 200 research questions that are necessary to guide the center's current work and potential projects. Aimed at addressing areas where current knowledge and means are insufficient for evaluating system safety, these research questions were identified through the exercise of CASMI's **Evaluation Framework**.

The Evaluation Framework outlines a mechanism to guide the assessment of system safety pre- and post-deployment. The Framework employs a decomposition approach, establishing the facts and issues of a system at the component level to then inform the system evaluation.

The Framework first progresses through a fact-gathering phase addressing the major areas in which system issues and research opportunities emerge:

- Data – Biases and other issues can arise throughout the data acquisition, cleaning, and normalization processes.
- Algorithm – The cycle of training and testing, the choice of specific algorithm, and the training and retraining dynamics each impact system performance and the nature of its application issues.
- Interaction – ML systems inherently interact with human users and the nature of that human interaction can determine much of the system's impact.

The Framework then progresses to evaluation informed by questions in the humanities and social sciences.

- Evaluation – Identifying and measuring the human impacts of an ML application must be done relative to the goals and values of the domain in which it is deployed.

The intent of CASMI's funded research at Northwestern and partner institutions is to further refine the roadmap or to address a key need in one of the Framework component areas. While the Evaluation Framework was a mechanism for developing the Roadmap, the Roadmap is a living document and an independent set of evolving research goals. As CASMI continues to uncover areas of opportunity, the Roadmap will be revised to prioritize research that advances the mission of operationalizing the safety of machine intelligence.

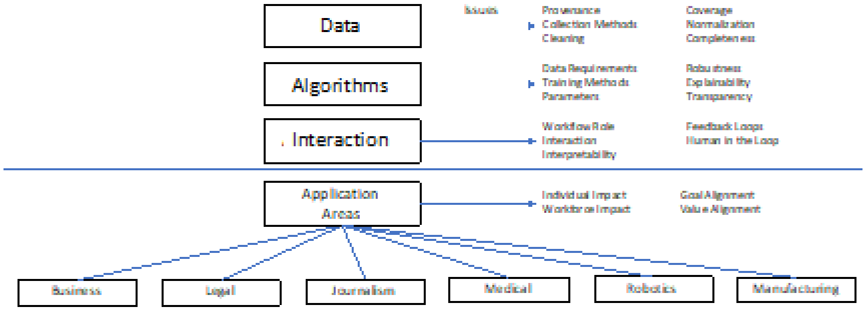
Table of Contents

Introduction

The goal of this Research Roadmap is to inform and prioritize work on the impact of Machine Learning (ML) systems on human health and safety. In defining this Roadmap, the focus is on both near-term and immediate harms (e.g., injury, economic impact, discrimination), as well as longer-term societal effects (e.g., the future of work, disruption of industries, job loss).

The precursor to this work – A Framework for the Design and Evaluation of Machine Learning Applications (or "Evaluation Framework") – segments the assessment of issues of the impact of ML into two primary phases: fact gathering, and evaluation of those facts with respect to specific harms. The motivation behind this segmentation was to develop a methodology that acknowledged the difference between the steps required to establish the core facts and those needed to evaluate the impact of the systems under different circumstances and goals. The Framework was designed to give some level of objective clarity to the process and to identify the core facts and features associated with any given application. This objective phase is then followed by what is often a more subjective process of evaluating those facts in the context of the goals and values they are designed to support and the negative impacts that might flow from them. (See "A Framework for the Design and Evaluation of Machine Learning Applications" for a full discussion of the Framework approach.)

The Evaluation Framework – by focusing on facts related to data, algorithms, and interactions – also provides a scaffold for identifying and understanding the work that needs to be done to operationalize the fact-finding process. It establishes the information needs for anyone evaluating a system and highlights whether those information needs can be met with current technologies and practice. In those areas where it is not possible to answer Framework questions (e.g., "Is a data set representative of a population?", "Which features were selected to guide training?", "Is an ongoing training process self-reinforcing?", "Does the framing of a solution skew the decision-making process?"), the gaps in our ability to answer those questions define the Research Roadmap.



Decoupling Facts from Evaluation

In parallel, the Framework provides a similar scaffolding for determining the research needs related to evaluation. While a more mature territory, questions of how to evaluate both the short- and long-term impact of digital technologies, how they are used in practice, and how different types of impact are traded off against each other remain open.

The Roadmap was developed through exercise of the Framework, which led to the development of a set of research issues that need to be addressed. These open questions were mapped into this Research Roadmap for each of the Framework component areas.

Finally, while the Evaluation Framework was used as a mechanism for the development of the Roadmap, the Roadmap stands as an independent set of research goals. Regardless of whether the Framework is used directly, the issues defined in this Roadmap are at the core of what we need to know if we are to understand the impact of the technologies of ML on human health and safety.

Research Roadmap

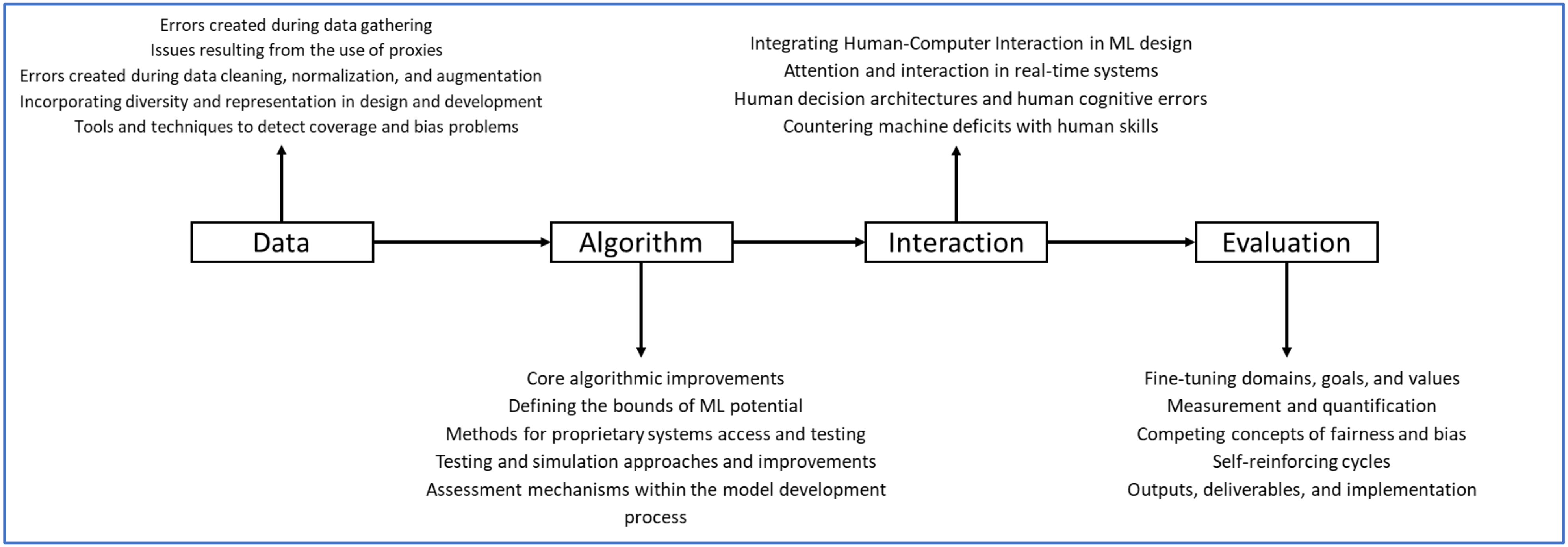
While the potential avenues of research in ML are vast, this effort is intended to focus on where this work will have value for the assessment of the human impact of ML systems. Thus, the development of the Evaluation Framework provided a structure to approach identification of gaps, key challenges, and under-addressed opportunities within the core component areas. Addressing these needs would allow for the operationalization of the Framework and assessment processes and move toward the assessment and mitigation of potential impact and harms.

The Research Roadmap is organized primarily by the research needs identified within the major component areas of the Framework: Data Gathering and Management, Algorithm Design and Application, Human-Computer Interaction, and Evaluation Approaches and Mechanisms.

The initial set of research themes and questions was compiled through a set of workshops and conversations with researchers in our primary areas of focus (data, algorithms, interaction, evaluation). Using the specific problems and issues that they surfaced as a guide, we developed a set of core research themes in each of these areas that we believe need to be explored. The key themes are explored in the sections that follow; Appendix A includes the list of the research questions within each theme.

In addition to the research opportunities within each of the discrete components, important research needs were surfaced that arch over component areas. These were primarily related to issues of policy and education, which may be less technological in nature but remain open research questions. A discussion of these is included in the Appendix A section “Additional Questions: Policy, Regulation and Education.”

While extensive, this Roadmap of research questions and opportunities is by no means exhaustive or complete. It is the starting point that encompasses an initial set of core themes and a set of high priority questions that need resolution today. As we move forward with the research plan, new questions and new research opportunities will arise. Using a metaphor from software development, each bug that is identified and fixed allows us to see additional ones that were hidden by the problems presented by the initial bug. And, as the field progresses at an ever-increasing pace, it will be incumbent upon us to reassess and reprioritize the Roadmap themes to ensure that ultimately the mission of evaluating and mitigating the human impact of ML systems is achieved.



Errors created during data gathering
Issues resulting from the use of proxies
Errors created during data cleaning, normalization, and augmentation
Incorporating diversity and representation in design and development
Tools and techniques to detect coverage and bias problems

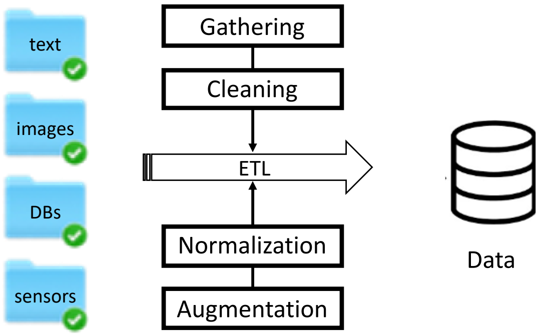
Integrating Human-Computer Interaction in ML design
Attention and interaction in real-time systems
Human decision architectures and human cognitive errors
Countering machine deficits with human skills

Core algorithmic improvements
Defining the bounds of ML potential
Methods for proprietary systems access and testing
Testing and simulation approaches and improvements
Assessment mechanisms within the model development process

Fine-tuning domains, goals, and values
Measurement and quantification
Competing concepts of fairness and bias
Self-reinforcing cycles
Outputs, deliverables, and implementation

Research Roadmap core themes by major component area.

Data Gathering and Management



Issues in Data Collection

The quality, completeness, and scope of data is foundational to any work in ML. Data issues cause a significant number of problems in applications supported by the models that the data enable. While it is tempting to tag these problems solely with the broad brush of *bias*, the issues are far more varied and are the result of errors at different stages of the data acquisition, cleaning, and normalization processes. While bias is a fundamental issue, the reality is that there are multiple sources of bias and it is one of many problems that flow from errors in data pipelines. Problems in data gathering, integrating multiple data sets, and the breadth of coverage of the data as it relates to target populations are all sources of issues downstream in the modeling process. The impact of data errors on resulting systems can be tremendous, and work needs to be done to develop the data management practices required to understand, design for, or evaluate these data sets comprehensively.

In our conversations with researchers and practitioners, we have identified five primary themes linked to the phases of the data pipeline that have significant impact on the design and engineering of ML systems:

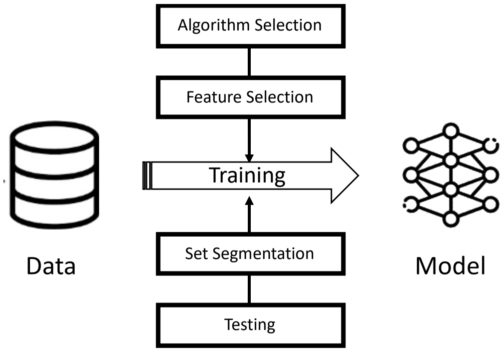
- Errors created during data gathering: The possibility of error or systematic bias in the data gathering phase of the ML process is known but not fully understood. Errors at this phase are often invisible to designers and developers doing later-stage data work and can cause problems that are hard to diagnose.
- Issues resulting from the use of proxies: Many important features that we want to predict would be difficult or impossible to measure directly or to represent in a quantified data set. To build systems that can be used to predict specific outcomes, developers are often forced to use related features to act as proxies. Research is needed to explore the ramifications of and mitigations for use of proxies.
- Errors created during data cleaning, normalization, and augmentation: As data are brought together and prepared for use in ML, they undergo a series of transformations. Empty cells may be filled, multiple tables joined, or new features inferred. Each of these has the potential to insert errors into the data. Research is required in how to identify, anticipate, and thus avoid these issues.
- Incorporating diversity and representation in design and development: Many problems in the development of systems, including those supported by ML, come from a lack of awareness of

target audiences, users, and those who might be impacted by the resulting systems. This is often exhibited at the data gathering/transformation stage of the work.
- Tools and techniques to detect coverage and bias problems: One long-range goal of this initiative is to develop best practices in data management that allow developers to avoid issues of uneven data coverage or inherent data bias. In the near term, there remains a need for tools that evaluate existing systems and their data sets for these issues.

Each of these themes shapes a research direction with multiple thrusts and research needs which are outlined in Appendix A, section Data Gathering and Management.

Algorithm Design and Application

The algorithm component of the Framework assesses the core technical apparatus of the ML system. As discussed in the Evaluation Framework, the key questions relate to the choices that are made in (1) designing and developing the learning algorithm and (2) tuning and testing the model. By probing where these questions can and cannot be answered, opportunity areas were surfaced where research needs to be done to understand, evaluate and guide the development of ML systems to prevent and mitigate harms and impact.



Issues in Model Building

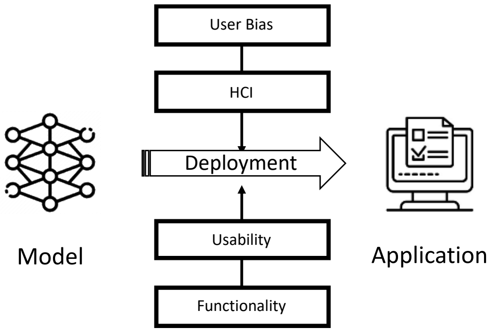
Five key themes arose amongst the opportunities for research in this space:

- Core algorithmic improvements: These research ideas include potential work to improve upon some of the most urgent known issues in algorithmic modeling that cause human harm and impact, as well as work to further classify the technical characteristics of ML models.
- Defining the bounds of ML potential: While much marketing material and popular press around ML describes its potential as limitless and ubiquitously applicable, these research topics look to specifically determine the technical bounds of ML models and their applicable use.
- Methods for proprietary systems access and testing: Given that most ML systems impacting individuals are built by private enterprises, a key issue in evaluating them is the proprietary nature of the technical workings of the systems. These research opportunities focus on developing methods to approach proprietary systems for impact evaluation.
- Testing and simulation approaches and improvements: These research opportunities focus on improving the methods and processes used to test and validate ML systems.
- Assessment mechanisms within the model development process: These opportunities focus on exploring specific mechanisms to evaluate the potential for impact or harm at identified appropriate points in the development process, in order to prevent and mitigate the issues earlier in the development lifecycle.

Each of these themes shapes a research direction with multiple thrusts and research needs which are outlined in Appendix A, section Algorithm Design and Application.

Human-Computer Interaction

No matter how precise or accurate models developed using ML are, in order to be useful, they must be embedded in a process workflow. At some point in that workflow, these embedded systems will be interacting with human users and the nature of that human interaction can determine much of the system's impact.



Issues in Application Integration

A system that recommends solutions is seen and used differently than one that provides critiques of user solutions. The former frames the solution that a user must respond to while the latter plays the role of devil's advocate to remind a user of aspects of a problem that they might have missed. These scenarios may utilize the same underlying technologies and models but have different outcomes. Likewise, a system that provides explanations in support of its answers could allow users to view those answers in a more mindful way and consider alternatives and hypotheticals while a "black box" system may be slavishly trusted or rejected without due consideration. And, in real-time situations, the shifting of control between a machine and its human users requires developers to understand how to switch context and control quickly and effectively.

Unfortunately, these interaction issues are often thought of as peripheral problems and can be ignored until developers are forced by ill-effects to focus on them. And, even when there is focus on human interaction in general, little attention is paid to the ways in which human cognitive skills and deficits impact the ways in which the results associated with a machine are interpreted and used in decision support. This is an area where work in Computer Science has fallen short and where much of the thinking that is required is found in the Social Sciences.

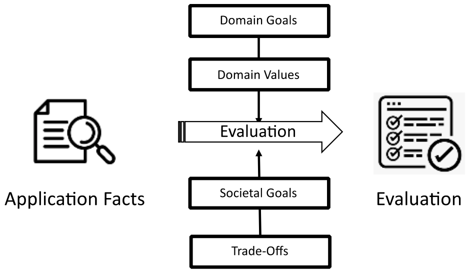
In conversations and workshops, we surfaced four themes that define the areas where there are gaps and open research problems that need to be addressed:

- Integrating Human-Computer Interaction in ML design: While Human-Computer Interaction is a field within Computer Science, the core ideas associated with attention, framing, and sequencing of information in the former rarely are transferred into the latter. Many of the open questions around this issue are focused on how to integrate models of HCI into ML development.
- Attention and interaction in real-time systems: As control is shifted from machine to user in real-time systems (e.g., self-driving cars), a persistent problem is the question of how to provide users with the context they need to make the right decisions in the moment.

- Human decision architectures and human cognitive errors: While much of the focus in ML research is on the data and the algorithms used, many problems and sources of negative impact come from a reliance on somewhat dated assumptions regarding human decision-making. Thus the problems that human decision-makers have because of confirmation bias, recency effects, anchoring effects, etc. can be amplified when working with systems intended to support decision-making processes.
- Countering machine deficits with human skills: Designing partnership ML systems and users is a process of balancing the skills and deficits of both sides of the collaboration. Deficits in the machine's ability to consider context, new data, or bias that has emerged can be compensated for with mindful human guidance. The question is how to best design for this human intervention.

Each of these themes shapes a research direction with multiple thrusts and research needs which are outlined in Appendix A, section Human-Computer Interaction.

Evaluation Approaches and Mechanisms



Evaluation Stage

The evaluation component of the Framework is focused on identifying and measuring the human impacts of a ML application, especially insofar as they relate to the goals and values of the specific domain in which it is deployed. By its nature, this component confronts challenging questions framed by both the humanities and the sciences, from the initial step of conceptualizing and identifying domains, goals, and values, down to the process of quantifying nebulous concepts like bias or fairness. Key themes that arise here arch over this whole process. Five key themes emerged:

- Fine-tuning domains, goals, and values: Developing precise methods for identifying and distinguishing domains, and then moving towards consensus goals and values for the domains under consideration.
- Measurement and quantification: What human impacts of ML can we currently measure? What impacts do we need to develop new measures for?
- Competing concepts of fairness and bias: Understanding the nature of harm and developing concrete measures of harms from bias, unfairness, etc.
- Self-reinforcing cycles: Identifying when self-reinforcing cycles are likely to occur, when they are problematic, and how to break out of them.
- Outputs, deliverables, and implementation: How should the outputs of this framework be presented? What are the most promising methods for framing and embedding an ongoing process of evaluation into existing institutions or processes?

Each of these themes shapes a research direction with multiple thrusts and research needs which are outlined in Appendix A, section Evaluation Approaches and Mechanisms.

Conclusion

The possibilities of Artificial Intelligence and ML are transformational. But just as there is potential for tremendous positive benefits, we have already seen that there is real risk for negative impact on human health and safety at both the individual and societal level.

As we move these technologies into use, we need to understand not only the possible ill-effects but their sources. Going beyond outcomes alone, we need to dig deeper into the root causes of bias, lack of coordination with human efforts, misalignment with personal and societal goals, and the errors that seem to plague ML systems in use today.

For each of the core phases of the development of ML-based systems – data, algorithmic choice, and interaction design – there are questions that need to be answered if we are going to develop operational models of safety in ML. Likewise, we need to develop methods for evaluation that consider human goals, values, and needs that can be used in both the evaluation of these systems and the construction of design principles for system development.

One of the more exciting aspects of this Roadmap and the open research opportunities it contains is how they were generated. They flow from a focused multi-disciplinary process that drew together experts in AI and ML, practitioners working on real-world deployment issues and researchers from the social sciences whose starting point is impact rather than the technologies themselves. While by no means exhaustive, they are the product of the kind of engaged conversations and sharing of perspectives we need to bridge the gap between examination of impact and technical issues that define and produce it.

We see this core methodology of cross-disciplinary work as the source of not just this initial Roadmap but also of an ongoing exploration of the problems we must address as the technologies of ML are brought into common use. The goal is to engage the broader social science community in informed examination of this impact with the insights then flowing back into the technical community to help guide the development of new systems and standards. We are striving for partnerships that help us to both define the problems and consider the entire ecosystem of conditions that give rise to them in crafting cross-disciplinary solutions.

The projects in this Roadmap will frame our efforts today. Even more important, the methodology has the potential to be a catalyst for the researchers in our network and beyond to seek out even more cross-disciplinary collaborations aimed at better understanding and guiding the impact of the systems we build.

We are eager to see the progress that will result from this work as passionate experts consider the needs of human health and safety in this burgeoning field. And we look forward to continuing to evolve the Roadmap with the next sets of research questions and opportunities that will push toward the development of ML systems that can promote and protect human health and safety.

Appendix A: Roadmap for Research Themes and Questions

This appendix enumerates the core research themes as well as specific research problems, opportunities and questions that have been identified as focus for the work of the Center for Advancing Safety of Machine Intelligence. Much of these were originally derived and developed from the workshops and conversations with researchers held as part of the Machine Learning Impact Initiative. Additional questions have been added as CASMI begins to work with its research network and advisory committees. As the research focus of CASMI is intended to progress in alignment with the needs in the field, this set of themes and questions will continue to evolve over time.

This document is organized by the core areas of focus from the components of the Evaluation Framework, with an additional section for issues that arch over the component structure.

Data Gathering and Management

The quality, completeness, and scope of data is a crucial part of any work in machine learning. Data issues cause a significant number of problems in applications supported by the models that the data enable. While it is tempting to tag these problems solely with the broad brush of *bias*, the issues are far more varied and are the result of errors at different stages of the data acquisition, cleaning, and normalization processes. In our conversations with researchers and practitioners, we have identified five primary themes linked to the phases of the data pipeline that have significant impact on the design and engineering of ML systems:

- (D01) Errors created during data gathering
- (D02) Issues resulting from the use of proxies
- (D03) Errors created during data cleaning, normalization, and augmentation
- (D04) Incorporating diversity and representation in design and development
- (D05) Tools and techniques to detect coverage and bias problems

Theme D01: Errors created during data gathering

While *human in the loop* is often seen as a desirable feature, individuals view the world in different ways, impacting data collection. Different individuals (and even the same individual at different times) will tag data, classify situations, and view features differently. This leads to uneven data sets and non-systematic errors. Methodologies are needed to manage these challenges and their impact on the data gathering processes.

- (Q1001) How can we develop systems for tracking, characterizing, and managing how human observations are biased and how that is thus mapped onto data gathering?
- (Q1002) Can we identify areas where crowd sourcing for data gathering might insert bias into data sets?

Data gathering interfaces can sometimes lead users in particular directions that shape the resulting data. Forms that require different levels of user input depending on previous answers can skew input toward paths of less resistance. Understanding the impact of interface design in this area would help to normalize and improve data gathering and ultimately data sets.

- (Q1003) How can data input and data gathering interfaces be designed to avoid introduction of errors and bias?

At a more general level, data scientists and machine learning engineers rarely directly manage the collection/gathering process and may not be aware of the issues that can arise during it. The result is models that are formally accurate but are flawed because of the data sourcing. Mechanisms that surface the possible problems and how to recognize them would go far in helping model developers recognize and then recover from errors that are made during data gathering.

- (Q1004) How can we create metrics for evaluation of data collection and develop best practices that can be used to direct data gathering and design?
- (Q1005) How can we use metrics for evaluating the validity of the data to rank the models that are subsequently generated?

The problem of quantification bias flows from an implicit bias in seeing numbers as inherently objective. As data is gathered and qualitative features are transformed into quantitative elements, guardrails are needed to ensure that the mapping is not simply a human “judgment call.”

- (Q1006) How should multidimensional and qualitative objectives be modeled to avoid quantification bias?

Theme D02: Issues resulting from the use of proxies

The targets for many uses of machine learning technologies are often subject to restrictions due to regulations, privacy issues, or proprietary elements related to the data. Even when there are no regulatory or privacy issues, some features may not exist in the form of machine usable data. To craft models that can be used to predict these features, developers will often use proxy data composed of features that that are more readily available. Health care costs are used as a proxy for health outcomes. Performance reports are used as a proxy for job fit. And click-through rates often stand in as proxies for interest or engagement.

While proxies can be useful, their use creates some striking issues and challenges with regard to their relationship with both the features they represent and the ground truth.

In current practice, there are no norms or standards for either selecting or evaluating the effectiveness of proxy features. Without such best practices that allow developers to view their proxy choices against the ground truth features that they are standing in for, it is possible to utilize proxies that do not track against reality.

- (Q1007) How can we develop methods for testing the reliability of proxy features against the features they represent and then expand those methods for use in the generation of proxies?
- (Q1008) How can we develop best practices based on the developed methods for testing reliability of proxies that can be applied easily within commercial environments?

As there is no single quantitative and measurable definition of many of the concepts that are often the ultimate targets of machine learning systems (for example, "successful employee" or "healthy person"), any proxy used will be imperfect; thus, methods for crafting and testing proxies need to be applied in these areas.

- (Q1009) Can we develop methods that allow us to identify when a quantitative definition of an objective is achievable, or the degree to which the proxy measures may misrepresent the intended objective?

Even when there are strong correlations between a target feature and a proxy, there are times when a proxy may be well correlated when judged against an entire population but skewed when applied to specific subcategories. Post treatment healthcare costs are well correlated with health outcomes in general but are far better correlated when patients have health insurance. As a result, they under-predict for patients who are under-insured or under-treated in general.

- (Q1010) Can we develop methods for evaluating the scope of the applicability of proxies that go beyond correlation to not only the ground truth features in general but their impact on prediction for different subpopulations?

The need to map restricted data onto unrestricted representations gives rise to the use of proxies. An alternative approach to this problem would be to develop methods for utilizing restricted data without violating any privacy issues.

- (Q1011) Rather than using correlating proxies, how can we develop methods for integrating protected data in ML models without violating privacy issues?

A consolidated set of best practices and mechanisms for addressing issues tied to proxy data would allow developers to identify proxies that are accurate and measure what is intended, rather than picking a proxy and then introducing another layer of subjective judgment to try to align it with ground truth. Metrics for measuring proxies would provide focus on the issue of the mapping between ground truth and the proxies selected.

- (Q1012) What tools can be developed to help uncover problems in bias associated with proxies for training data?
- (Q1013) What mechanisms would allow for the establishment of bounds on the applicability of proxy metrics to represent the uncertainty and potential error of the proxy?
- (Q1014) Is it possible to test the reliability of proxy measures as indicators of the training targets?

Identifying practical approaches, available mechanisms, and tools for the development community could influence proxy utilization in practice and the ultimate impact.

- (Q1015) What steps are needed to develop courses for data engineers on the issues of proxies, the possible problems they might introduce, and best practices for selecting and validating them?

Feedback flowing from existing applications could be better utilized to understand the impact of the proxy data and to identify approaches to aid in the selection of proxies.

- (Q1016) Can feedback from successful or unsuccessful achievement of outcomes be used to improve and to identify data proxy approaches or interpretation?

Tools for identifying proxies by testing them against each other would improve selection of proxies and the resulting systems.

- (Q1017) How can we assess the sensitivity of predictions to different choices of proxies?

Theme D03: Errors created during data cleaning, normalization, and augmentation

There is a disconnect between data engineers who are collecting and integrating data into usable sets and the data scientists or machine learning engineers who are using that data to develop models. The cleaning, normalization, and augmentation process is viewed as separate from the modeling itself. As a result, machine learning engineers rarely have any sense of the source of their data or what has been done to it. Likewise, data engineers rarely have a sense of the implications of decisions that they have made while pulling their data into a coherent whole. This leads to situations where data layer issues are not recognized, and therefore not mitigated.

A major component of this problem is the lack of mechanisms for tracking, representing, and communicating the history of the changes that have been made to a data set. Explicit representation and sharing of data provenance would provide machine learning engineers with a better understanding of the source of their data and possible issues with it. Current techniques for data development do not include explicit representations of data provenance. Tracking and representing all changes would support methods for using the representation in learning and in communicating results.

- (Q1018) How can we develop the vocabulary or language that allows us to represent the actions and states that define the provenance of data?
- (Q1019) Given a language to describe the steps associated with data management and transformation, can we develop tools to track and organize information about the processes used in data development?
    - Tools to record versioning, data lineage, and basic provenance of all data sets.
    - Tools and techniques to track and to record the steps taken to join data sets together data before training.
    - Tools to track processes aimed at data enhancement and augmentation.

While problems within data that lead to errors can be characterized in the abstract, more refined characterization of issues associated with different kinds of data and different gathering techniques would support the development of better methodologies to avoid these issues.

- (Q1020) Can we develop a taxonomy of approaches to error detection related to different types of data (images, text, audio, video, etc.) and how they were gathered?

The generation of synthetic data can introduce error and bias. With a clear taxonomy of causes and sources of error, some of these problems could be avoided. This taxonomy may also be used as the basis of an approach to mitigate existing bias issues. Work is needed to determine the capability and applicability of this method.

- (Q1021) Given a taxonomy of known sources of error, how can that information be used to guide the generation of synthetic data sets?

One of the more pressing issues in data cleaning is the problem of developer bias. Developers often make decisions about the data they are managing that includes removing rows, merging different data sets, or having one set of features supersede others. This ends up transferring the developer's biases into the data itself.

- (Q1022) How can we help developers by teaching them how to stop their value judgments seeping into data development?

Current approaches to equity in data tend to be ad hoc and may be susceptible to the tester's own bias. Standardized approaches and mechanisms for the examination of proxies could help to identify equity issues and correct for the bias in the tester.

- (Q1023) What approaches can be developed to reliably test the equity across user classes in candidate proxy selections?

Many data sets are simply too small for machine learning to be applied responsibly and accurately. To deal with this problem, we need techniques for augmentation and the generation of reliable and robust synthetic data sets, as well as tools to assess when and how it is appropriate to do so.

- (Q1024) How can we best address low-*N* conditions (e.g., when can we impute or synthesize data instead)?

Techniques for incentivizing data collection and validation (versus the system) can provide the basis for a greater focus on data management.

- (Q1025) Is it possible to place a premium on training data that is neutral relative to the predictions and the feedback loops that may otherwise reinforce the training?

Rather than waiting for post-deployment testing, simulation can be used as a tool for testing coverage (equity), correspondence with existing practice (error), and equal application (bias).

- (Q1026) What role can large-scale simulation play in the determination of equity and testing for error and bias?

Data use and reuse remains ad hoc. Best practice around how data has been gathered and augmented and what functional pressure has been put on it would provide metrics for when (and when not) to reuse it.

- (Q1027) What best practices can be developed around data reuse and improvement?

Theme D04: Incorporating diversity and representation in design and development

Stepping away from engineering-only approaches, techniques from the social sciences could provide greater leverage in the design and testing of data layer equity. For any given system, engaging the appropriate stakeholders and incorporating an understanding of their needs changes the nature of the goals that drive the design. Drawing in diverse stakeholders may help to incorporate and balance the needs of impacted groups and potentially avoid issues common in data gathering.

- (Q1028) What multi-disciplinary approaches can be brought to bear to assess reliability and equity?
- (Q1029) How can broader sets of stakeholders inform the design and evaluation of data sets?
- (Q1030) How can broader sets of diverse stakeholders help in data gathering and design?

The nature of bias, fairness, and equity shift depending on the stakeholders and cultural context of an application. Providing better cultural diversity can expand the set of issues that can be tested for in general and specifically in language models.

- (Q1031) What opportunities are there for exploring cross-cultural and multi-theoretical comparison of language model biases?

Theme D05: Tools and techniques to detect coverage and bias problems

One of the more difficult issues that we face is the language we use in describing issues such as bias in data sets. The term is often used to describe the outcomes associated with the use of a data set or structure but not the sources. At the data layer, we understand sample, prejudicial, and measurement bias and tend to stay at the level of these three sources of error rather dive deeper into causes. Sample bias, for example, can come from faulty survey methods, sensor failures, crowd sourcing, or lack of diversity in data sources.

A standard, consistent, and agreed upon model of the sources of bias at the data layer and techniques for anticipating and avoiding them would be a key step in building data sets that are closer to a valid reflection of ground truth. A standard representation designed to annotate known sources of bias and to propagate that information through the ML pipeline would allow downstream systems to adapt the data as needed.

- (Q1032) How can we develop a vocabulary of bias that goes beyond the standard ideas of sample, prejudicial, and measurement bias to a more detailed characterization of causes?
- (Q1033) How can we map a vocabulary of bias onto best practices for data engineers?
- (Q1034) How can a vocabulary of bias be used to create a data bias assessment process that provides a mechanism to tag data sets regarding potential bias?
- (Q1035) How can the resulting tagging be integrated into data provenance meta data?

For data sets that are hidden (e.g., proprietary), their impact can be seen only once a model is built. Upstream problems of coverage and correctness need to be exposed to isolate potential issues before they occur. Data with personal information needs to be protected but also needs to be evaluated. The techniques of differential privacy used in medical settings and federated data systems should be explored for adaptation to the evaluations required for ML systems.

- (Q1036) What steps can be taken to validate corpora without direct examination?
- (Q1037) To what extent can ideas from differential privacy be used to help protect data models while still auditing them for not having bias?

One of the greatest sources of bias from data is the use of data sets that do not provide coverage of examples that match the distribution in the real world of the target populations. Methods are needed for testing for coverage against populations.

- (Q1038) Independent of how data is sourced, how can we develop mechanisms to audit data distribution and coverage as a first step before model considerations and training?
- (Q1039) What methods can be developed to construct representative datasets: building datasets that include more than just the successful or positive cases to represent the actual instances in the full population and failures or negative cases?
- (Q1040) What tools can be developed for data profiling to identify imbalanced representations?

Data augmentation is driven by the needs of the systems that are being built. As a result, augmentation techniques are applied to fit those needs, regardless of whether they maintain the integrity of the data set. Augmentation tools and techniques are needed to assure the integrity of the data and thus the resulting model.

- (Q1041) Can we develop data augmentation techniques that are guaranteed to be truth preserving?
- (Q1042) Can we develop measures and metrics for data augmentation that allow us to assess the risks of adding bias or noise to the data?

Metrics are needed that appropriately define and evaluate the features of good data sets.

- (Q1043) What is the data version of a 'gold standard' that would ensure that data is appropriate for the task and has the right provenance?

Tools that have clear and operational metrics for coverage and diversity could move the process away from an ad hoc evaluation of data.

- (Q1044) What tools can be developed to assess the suitability, coverage and diversity of data sets identified for training and use?

Mechanisms need to be developed that may identify whether a dataset is comprehensive of the full process or may be narrow in scope such that it reflects a distorted view of the skew of the population.

- (Q1045) What mechanisms would engender more focus on understanding the data-generating process, not just the data?

Algorithm Design and Application

The algorithm component of the Framework assesses the core technical apparatus of machine learning systems. As discussed in the Evaluation Framework, the key questions relate to the choices that are made in (1) designing and developing the learning algorithm and (2) tuning and testing the model. By probing where these questions can and cannot be answered, opportunity areas surfaced where research needs to be done to understand, evaluate, and guide the development of ML systems to prevent and mitigate harms and impact. Five key themes arose amongst the opportunities for research in this space:

- (A01) Core algorithmic improvements
- (A02) Defining the bounds of ML potential
- (A03) Methods for proprietary systems access and testing
- (A04) Testing and simulation approaches and improvements
- (A05) Assessment mechanisms within the model development process

Each of these themes follows with a list of the concepts that were identified as important next step research opportunities.

Theme A01: Core algorithmic improvements

Developers often switch between approaches based on the precision/recall numbers of models without regard to the requirements and tradeoffs. They have little guidance as to what is lost or gained by switching between approaches and how they alter outcomes and impact.

- (Q1046) How can we establish and disseminate a compendium of the fundamental tradeoffs between algorithmic choices?
- (Q1047) How do we map and communicate the core science of the tradeoffs of algorithmic choices onto clear best practices for engineers?

We currently have no best practices for avoiding inappropriate proxy and/or narrow reward functions as we build models.

- (Q1048) Can we develop methods for designing reward functions that would enable responsible reinforcement learning?
- (Q1049) Can we develop methods and best practices for choosing proxies for otherwise untagged data?

We do not have a clear understanding of how shifts in distribution in our training sets impact outcomes. As a result, engineering decisions concerning data sets often have unforeseen impact on results.

- (Q1050) Can we determine the implications of prediction changes that result from shifting models?

The unsupervised nature of language models makes them especially prone to bias based on the training materials. Manual methods of testing for that bias are insufficient. The current state of the art for testing language models utilizes focused manual queries that examine known biases. Automated techniques that can be applied more broadly could uncover bias and unforeseen term relationships. Development of automated methods would allow developers to discover and potentially to mitigate bias more effectively.

- (Q1051) Can we develop methods for identifying recognized bias in language models at scale and regardless of context?
- (Q1052) Can we develop methods for removing bias from these models as they are applied?
- (Q1053) How can we uncover unknown/unexpected biases or unwanted relationships between terms in language models?
- (Q1054) Can we develop methods for decoupling inappropriate or undesired links between terms that were built during training of large language models?

Research is needed in how to determine the contexts in which the introduction of a ML system may cause self-reinforcement and scenario fulfillment and methods to be able to determine to what degree that would likely occur. When would the presence of a human, or their observation, be likely to affect the phenomenon that’s being predicted? This could result in developing design mechanisms such that the model can account for the observation as a fundamental part of the learning process and take that into account when making predictions and modeling the underlying process.

- (Q1055) Can we develop methods for examining feedback loops, self-reinforcement, and scenario fulfillment in ML models?

Ensemble methods could utilize multiple models to verify or to balance each other.

- (Q1056) Are there multi-model mechanisms that can correct for the bias exhibited in single model systems?

Theme A02: Defining the bounds of ML potential

ML systems have difficulty accounting for new data that is outside of the initial data set for any reason (because it was outside of the collection process, is rare, or may be yet to happen/a future scenario).

- (Q1057) Can we develop methods for recognizing *black swan* or *broken leg*[1] conditions that impact the predictive effectiveness of ML predictions?
- (Q1058) Can we develop methods for incorporating unknown conditions into existing models that do not require complete retraining?

In the universal function approximation theorem, the idea is that with enough capacity and enough data a model can recover any input to output function – but what happens when the classification induces an underlying data shift (e.g., there's some feedback loop between data and output)? Research is needed

[1] When an unexpected new feature (“Bob broke his leg last night.”) impacts predictions based on historical data (“Bob buys coffee at Starbucks each morning.”)

to determine how and to what extent technical methods can account for interaction in the ongoing learning process.

- (Q1059) Can we develop methods for recognizing conditions in which feedback loops are shaping outputs?
- (Q1060) If feedback loops are identified, can we develop techniques for mitigating their impact on models?
- (Q1061) Can ML techniques be developed with interaction loops in mind so that the systems themselves can monitor and compensate for their impact?

There are opportunities to utilize problematic or biased prediction systems as diagnostic agents to further investigate the potential systemic or historic issues in a particular context or process which then may need to be addressed with holistic methods.

- (Q1062) Can we develop a set of tests for models aimed at identifying bias in the world rather than bias in the data or the systems themselves?
- (Q1063) Can we design ML systems with a diagnostic focus rather than a focus for action?
- (Q1064) Can we use ML systems to uncover bias and inequality in systems by learning from the data associated with them?

Theme A03: Methods for proprietary systems access and testing

Many of the emerging models in use today are based on data found behind firewalls designed to protect personal information, making them hard to evaluate. Likewise, the models produced using this data are not open to testing if those tests have the potential to uncover proprietary information. This means that there are severe limits on our ability to unpack them for evaluation.

- (Q1065) Can we develop mechanisms for providing test access to models without compromising privacy so that we can better understand and respond to them?
- (Q1066) Can we extend proprietary system testing mechanisms to develop testing harnesses that can provide the ability to evaluate these models without compromising either personal or proprietary information?
- (Q1067) Can we develop mechanisms for evaluation of proprietary models to probe them without compromising Intellectual Property?
- (Q1068) What approaches, methodologies, and mechanisms would allow for testing black boxes in general?

Theme A04: Testing and simulation approaches and improvements

Many systems are tested against subsets of historic data, simulations, or a developer's notion of the world. But these are crafted to support the work and sometimes fail to test the outputs. With better simulators and testing harnesses, problems can be identified upstream, avoiding the common outcome of ML systems that have significantly poorer results in production than were demonstrated in testing. We need to investigate the development of production sandbox environments where researchers can put models in front of users to study training/deployment mismatch.

- (Q1069) How can we develop testing and training paradigms that more closely mimic the real-world deployment environments that will be encountered when a model is deployed?

- (Q1070) Can we develop a sandbox or simulation environment that would allow researchers to deploy their models in a simulated real-world environment where it is interacting with users without having to be production deployed?

Theme A05: Assessment mechanisms within the model development process

Establishment of test/training distributions tend to be manual in nature. With tools and/or automation, the process can be streamlined and normalized. The goal would be the development of tools that incorporate best practices for testing example distributions, resulting in more predictable outcomes. Establishing a best practice and rigorous mechanisms regarding example distributions could relieve the decision-making pressure from developers and appropriately pause the launch of systems that do not have adequate distributions.

- (Q1071) Can we develop automated detection mechanisms to identify when test distributions differ from training distributions?
- (Q1072) How can developers know when it is acceptable to deploy or use a system if it has distribution errors?
- (Q1073) Can we develop tools that automatically test example distributions against real-world distributions?

In model development, most engineers have limited information about either the source of the data that they are using or how the model they are building will be used in practice. This limits their ability to test models against anything that scopes beyond the data that they have in hand and the specific performance targets that they have been given. In particular, they do not have access to domain-level goals and values that might be impacted by the outputs of their models. Providing this awareness would help to limit the problem of “unintended consequences” that are so often cited when problems in model application occur.

- (Q1074) Can we develop a language and mechanisms for defining domain-level goal/value characterizations that can be used in the development of requirements for ML engineers?
- (Q1075) Can we develop a process model that scopes across domains to provide guidance for developers that goes beyond the current ad hoc approaches at the domain level?
- (Q1076) Are there domain-agnostic processes that can be used in evaluative pipelines throughout the workflow for bias tests and checks?

Depending on when a system is checked, determination of bias and fairness might vary. When is the right time to begin checking that doesn’t undercut the development process or simply wait until the very end of the process?

- (Q1077) At what stage(s) can and should the system be checked for biases?
- (Q1078) How early in development or training would be viable to perform system checks for bias?

Explanatory AI methods only “explain” a specific part, representation, or type of output. How do we develop the appropriate tasks that demonstrate when explanations are useful (or not)? For opaque models, in particular deep learning models, explanations of their operation could be used to identify issues and to debug them.

- (Q1079) Can we define rigorous definitions for explanation at different stages of the process of developing, debugging, and deploying models?
- (Q1080) Can we generate explanations aimed at developers that can be used to better understand and debug models?
- (Q1081) Can we generate explanations aimed at supporting greater cooperation between systems and their users?
- (Q1082) Can we generate explanations in support of auditing of systems from the perspective of responsibility and regulation?
- (Q1083) Can we generate explanations in support of stakeholders who are not users but are impacted by the outcomes associated with specific systems?
- (Q1084) How can we best evaluate explanations for their ability to be helpful to users and algorithms?
- (Q1085) How do we extend methods used in the development of explanations for neural models to other types of data and systems such as those built using time-series data or collaborative filtering models?
- (Q1086) Can we develop explanatory methods that are transferable between multiple problems and domains?

Human-Computer Interaction

No matter how precise or accurate ML models are, to be useful, they must be embedded in a process workflow. At some point in that workflow, these embedded systems will be interacting with human users and the nature of that human interaction can determine much of the system's impact.

This is an area where work in Computer Science has fallen short and where much of the thinking that is required is found in the Social Sciences, in work in human decision-making architectures. Even within the field of Human Computer Interaction (HCI), the primary focus of the work has been on usability rather than ultimate impact. The result has been a gap between developers, the goals of systems that they are crafting, and the skills and limitations of the people who will use them.

We have surfaced four themes that define the areas where there are gaps and open research problems that need to be addressed as we consider how to evaluate and to design the interactions between ML systems and the humans who use them:

- (I01) Integrating Human-Computer Interaction in ML Design
- (I02) Attention and interaction in real-time systems
- (I03) Human decision architectures and human cognitive errors
- (I04) Countering machine deficits with human skills

Each of these themes follows with a set of the research questions that were identified as important next step research opportunities.

Theme I01: Integrating Human-Computer Interaction in ML design

If we are to balance machine and human skills, we need to bring work in HCI, human factors, and behavioral economics into the development of AI/ML systems. We need to recognize that ML applications function within a larger process ecosystem that requires connection to and accommodation of human reasoning styles. Systems that are extensions of human thought rather than being at odds with it could be integrated with workflow for better systemic outcomes.

- (Q1087) What is needed to bring HCI and UX researchers to the table in ML systems design?
- (Q1088) How can the body of work on human factors be better incorporated and referenced in the design process for ML applications?
- (Q1089) How can we adapt the work on how people interact with each other to improve the outcomes as they interact with intelligent systems?
- (Q1090) How can we adapt the work on how people interact with other devices to improve the outcomes as they interact with intelligent systems?
- (Q1091) How can the design of applications be approached as extensions of, or tools for, the mental and psychological processes of human decision making?

The ability to effectively use tools depends on users understanding their functionality. Improving the communication will help improve these models and, in turn, improve outcomes.

- (Q1092) How can issues with data or models be communicated to users to help them understand and incorporate machine suggestions more mindfully?
- (Q1093) How can we understand how users develop mental models of a system?

Most users have a weak understanding of statistics so information that takes the form of statistical assessments are often misunderstood. Methods to better contextualize and communicate information about certainty and likelihoods would better support user decision-making. For example, without an understanding of the uncertainty issues associated with a system it is difficult for users to assess how to interpret results. Mechanisms to help them do so would go far regarding the appropriate utilization of the results flowing from models.

- (Q1094) How can ML predictions be presented in a way that goes beyond statistics?
- (Q1095) Is it possible to communicate uncertainty in a way that is understandable to users, even those who may have difficulty with numeracy?

Users can be primed and have their decisions skewed simply by the way in which a system suggests possible answers. Mechanisms for muting this priming would support more thoughtful decision making in partnership with systems.

- (Q1096) Is it possible to present users with predictions without priming them or biasing their judgments?

While significant focus has been placed on developing mechanisms for explainability in AI systems, research is needed to further understand how humans utilize explanations and how different explanation constructions could impact the decision-making process and its effectiveness.

- (Q1097) What sorts of explanations are the most useful in helping support human decision making?

Holistic, sociotechnical framings of the full context of systems could further their effective and appropriate use and promote understanding of what and why they do things. Improving this contextual understanding will help improve these models and, in turn, improve outcomes.

- (Q1098) How can we develop human factors and sociotechnical approaches to system development that reflect the entirety of the system, not just the technical/algorithm components?

Theme I02: Attention and interaction in real-time systems

By understanding where and how human decision making is limited, linked to impactful outcomes, or too easily directed, we can prioritize and define application development in appropriate contexts and bounds. One area that is particularly problematic is that of attention. In the absence of prompts or processes that maintain attention, user focus can waver, which can lead to problematic outcomes in interacting with a system that is assuming full attention and partnership.

- (Q1099) Given the constraints of human attention and focus, can we determine reasonable tasks that can be automated or partially automated?
- (Q1100) Can we develop principles of attention monitoring and prompting that reinforce focus and situational awareness?

Given that one of the most crucial moments in human/computer hybrid systems is the hand-off from one to the other, a mindful approach to deciding and controlling that hand-off could decrease the likelihood of negative outcomes.

- (Q1101) Is it possible to determine when the machine should cede control to humans and when the machine needs to assert itself and take control back?
- (Q1102) Can we develop methods that allow a system to track the conditions that will lead to the need to change control and prior to that step, engage users to give them time to establish situational awareness?
- (Q1103) Can we develop methods that allow systems to track their own users and engage them even when the machine remains in control?
  - How do we artificially keep people engaged?
  - What methods and mechanisms can be developed to help humans avoid distraction?
  - How do we design activities that help users maintain situational awareness?
- (Q1104) Can we map these developed methods for user engagement into evaluation metrics and developer best practices for real-time systems?

The skills required for human-in-the-loop may be different than those used for the task in general. Skills in the latter may not transfer over to the former suggesting that development of specific training methods is needed.

- (Q1105) Is it possible to train users to more effectively use real-time systems that require handoffs?

A model of how much a user is attending to a situation would go far in the decision-making process of when to hand off control and how to set the context for that hand-off. Given that a substantial number of issues with real-time systems are related to hand-offs, a model that explicitly manages them could correct a wide range of problems.

- (Q1106) How should a system that hands off control represent attentiveness on the part of the user?

- (Q1107) Can collaborative tasks be broken down into modules that have more explicit hand-offs?
- (Q1108) Can we develop models in which users maintain control while the intelligent system performs subtasks automatically?

Developing new methods for handing off control can avoid issues where human users are asked to take control without context. These may be able to be informed by existing effective elements of human-human hand-off processes.

- (Q1109) What are the different types of human-human hand-offs and what can be learned to apply to the problem of human-machine hand-offs?

Mapping out the areas where control should remain with the human user, either based on domain or machine skills, could reduce the likelihood of problems.

- (Q1110) What domains and contexts are ready for a human-machine handoff, and where would it be inappropriate?

Theme I03: Human decision architectures and human cognitive errors

Systems have models at their core that are assumed to utilize relevant features in their operation. But many utilize features that come from a broader context. Users who do not have access to the full range of features and context will be challenged to understand or interpret the results.

- (Q1111) Can we design systems that represent and communicate their "contexts" explicitly?
- (Q1112) Can we represent the different contexts that impact decisions outside of the standard models of the core factors used in learning?

Often systems are used for functions for which they were not intended, sometimes to ill-effect. Mechanisms to identify the potential for a system to be used in other contexts and considering those contexts in the design and development process could block unintended consequences based on out of design usage.

- (Q1113) How can we develop mechanisms to identify the likelihood of a system being applied in a context outside of its original design?
- (Q1114) Can we determine the areas in which individuals may be likely to use technology in emergent or contraindicated ways that may cause harm?

Understanding how human reasoning works can provide system builders with guidance as to the areas in which they may be prone to errors while working with a system.

- (Q1115) What aspects of human decision architecture make individuals prone to error or misunderstanding as they interact with machines?

Friction in a system is a point where the design is intentionally more challenging or cumbersome to cause the user to have to engage with the process. This can maintain a space for human judgment and decision making to contribute to the ultimate outcome.

- (Q1116) What are methods for and what is the effectiveness of the introduction of friction to processes, to help address that the human user doesn't unduly trust the system?

These methods may identify the scenarios where humans are being primed to look for a subjective phenomenon, such that if they are told to go to a place and look for it, they are more likely to find it. This may also be where knowledge of predictions might affect human behavior.

- (Q1117) What metrics and methods can be developed to identify the factors and contexts where an AI model is likely to contribute to a self-reinforcing cycle or scenario fulfillment?
- (Q1118) How can we identify applications where the use of a prediction will lead to a feedback loop?

Humans tend to "use up" the slack in systems if there is no immediate impact on them. As a result, they sometimes have trouble keeping a mindful attitude and remain aware of what a system is doing and when they need to intervene.

- (Q1119) How can we accommodate the tendency of humans to "use up" additional safety that they're given, e.g., when they know they have airbags, seatbelts, and lane-keeping, they drive more recklessly?

One of the side effects of deploying automated systems to take on tasks previously performed by humans is the atrophy of human skills. Determining the appropriate training mechanisms to counter this degradation of skills as these systems are used is critical.

- (Q1120) Can we identify when retraining is necessary, given that there may be cases where machines can fully automate the task and take it out of the hands of humans?
- (Q1121) If systems become more responsible for certain tasks and that leads to atrophy of skills in humans, how do we train them differently than we do today to mitigate or compensate for that skill atrophy?

Theme I04: Countering machine deficits with human skills

Designing partnerships between ML systems and users is a process of balancing the skills and deficits of both sides of the collaboration. Deficits in the machine's ability to consider context, new data, or bias can be compensated for with mindful human guidance. The question is how to best design for this human intervention.

The notion of "human in the loop" may be a crutch rather than a tool. Identifying where it can be effective and where it is likely to be abused could establish new rules for deployment.

- (Q1122) Is a "human in the loop" helpful in correcting machine errors?
- (Q1123) Is it possible to develop a training process and protocol for individuals who use predictive algorithms in fraught or potentially dangerous contexts?
- (Q1124) How can we ensure that users have appropriate mental models of technologies in use and determine the extent to which the appropriate mental model may be able to mitigate potential harms?

There are times when a system as it is developed may be at odds with the goals and values of its domain, in particular when those goals and values may be changing and when the system is likely to reinforce or sustain a value system based on a historical context.

- (Q1125) Are there mechanisms that would foster a pause on certain interventions that lead to reinforcement cycles where society is rethinking the questions, goals, and values in that context?

One approach to bias in a system is to develop mechanisms and methods that would allow users to recognize, interpret, and mitigate the harms that could arise from the systemic or historic issues that exist. We need to define the technical and procedural mechanisms that would allow human guidance and correction of a system, and to further develop those in the appropriate and necessary contexts.

- (Q1126) How can interactions with users be used to counter systemic bias in existing systems?
- (Q1127) Are there mechanisms that could be developed to provide ongoing correction of the predictions or assessments of an ML system based on ongoing human feedback?

Evaluation Approaches and Mechanisms

The evaluation component of the Framework is focused on identifying and measuring the human impacts of a machine learning application, especially insofar as they relate to the goals and values of the specific domain in which it is deployed. By its nature, this component confronts challenging questions framed by both the humanities and the sciences, from the initial step of conceptualizing and identifying domains, goals, and values, down to the process of quantifying nebulous concepts like bias or fairness. Key themes that arise here arch over this whole process.

Five key themes emerged:

- (E01) Fine-tuning domains, goals, and values
- (E02) Measurement and quantification
- (E03) Competing concepts of fairness and bias
- (E04) Self-reinforcing cycles
- (E05) Outputs, deliverables, and implementation

Each of these themes follows with a set of the concepts and open questions that were identified as important next step research opportunities.

Theme E01: Fine-tuning domains, goals, and values

One of the pressing issues in ML is the disconnect between the language of goals, values, and ethics and the language of constraints, requirements, and features. Defining how this can be translated in practice is necessary for both the evaluation of existing applications and the development of requirements for future development.

- (Q1128) How can we frame goals and values using the language of constraints, requirements, and features that can be used to define applications for engineering?
- (Q1129) How can we identify when a system is not aligned with the domain in which it is deployed?

Goals and values may imply a focus on positive expectations. Going beyond this to consider explicit domain and societal constraints in the evaluation may help to articulate what is not acceptable or specific minimum acceptance criteria in the system and outcomes.

- (Q1130) Can we expand the idea of domain *goals and values* to include the notion of societal *constraints* as part of the evaluation process?
- (Q1131) What are the possible points of connection and points of friction between domain-level goals, individuals, and society?

Some goals and values are culturally bound. Explicitly identifying commonalities and differences that range across cultures would serve as guides to the ethical development of machine learning.

- (Q1132) How can we describe the cultural differences in goals and values?
- (Q1133) How can we map cultural differences in goals and values onto a language of design and evaluation?

Our goal is to map evaluation back onto upstream design principles that, if applied, align systems with the appropriate goals and values. With this step, we can provide developers with the tools they need to focus the design work on the goals and values that will be part of the evaluation downstream.

Rather than evaluating the consistency of goals and values after a system is built or deployed, research is needed to establish and assess design processes and mechanisms that could avoid or prevent misalignment issues.

- (Q1134) Can we incorporate downstream impacts as part of the fundamental validity of the system itself (e.g., teaching to the test, consequential validity)?
- (Q1135) What design processes and mechanisms would ensure upfront alignment or consensus on goals and values, then an examination of data (and proxies) for fit to that alignment?

Theme E02: Measurement and quantification

A major problem in the adoption of responsible methods for AI development is the lack of a quantifiable set of metrics that designers and engineers can use in the development of systems. A mapping of the more qualitative language of ethics onto quantitative metrics for engineering would further the integration of these goals and values into machine learning projects and into decision-making processes; it would also help to articulate and measure the resulting impacts and harms of ML systems. This is essential if we are driving towards a set of data-driven best practices, so that the human impacts are not “just anecdotal.”

- (Q1136) What is the translation of values and goals into concrete, actionable, quantifiable impacts?
- (Q1137) How do we translate inherently qualitative goals onto quantitative metrics?
- (Q1138) What kinds of impacts count as harms and how do we measure them?
- (Q1139) What specific kinds of harms arise as issues in ML systems?
- (Q1140) How can we identify and measure those harms?
- (Q1141) How can we quantify the impacts of various design choices and mitigation strategies?

In some domains and contexts, it is possible to identify the ground truth for the occurrence of some features, and then use that to de-bias algorithms and generate missing data reliably. This is probably not possible for other contexts. For example, since data about crime is thoroughly veiled behind many layers of social construction, interpretation, and judgment, crime data may not have a robust ground truth comparator.

- (Q1142) What applications or domains are defined such that “ground truth,” even with incomplete data, is a robust construct?
- (Q1143) In those domains where ground truth is less well defined, how do we map performance onto measures of impact?

- (Q1144) Are there mechanisms to better assess or de-bias ground truth than the current proxies in complex cases?

ML systems are often deployed in areas where they are replacing human reasoning with machine classification and prediction. We need to be able to understand how the introduction of prediction mechanisms are different than those being made by existing processes.

- (Q1145) How do we measure the before and after impacts of ML as it is deployed?
- (Q1146) What are the evaluation mechanisms that assess how and to what degree algorithms augment the way (non-algorithmic) predictions are already being made?

The communication gap between the technical and non-technical staff in companies leads to a misalignment of how to evaluate systems and the goals that they should achieve. A bridge between these stakeholders would help developers to better construct systems that are designed around both business and social goals.

- (Q1147) Can we establish processes that better enable designers, developers, and data scientists to effectively communicate with stakeholders about appropriate measurement, evaluations, and goal/value decisions?

Research is needed to understand if it's possible to identify where there may be fundamental misalignment of the properties of machine learning with a domain's goals and values and to devise appropriate protocols for further ML work that may be done in that domain.

- (Q1148) Can we identify areas where ML use is at odds with a domain's goals and values, or not likely to be successful, and an approach for justification for doing research in those areas?

Operationalizing ethics involves not only establishing the rules but also establishing how they comport with human behavior. This can further alignment between rules and how they fit into the decision-making process.

- (Q1149) How can functional definitions and measurements of ethical goals be developed with human psychology in mind?

Theme E03: Competing concepts of fairness and bias

All learned systems, both machine and human, have some level of bias. We need to develop a model of how to define it and its impact rigorously. Defining the different metrics of bias and impact (both positive and negative) would provide a framework for deciding which can be universally scoped and which are more aligned with cultural or societal norms.

- (Q1150) What are the metrics of bias itself and how is it identified and defined?
- (Q1151) Are the metrics of bias universal or how would they need to be modified or redefined in different contexts, geographies, or constituencies?

Developing metrics that go beyond the binary and capture the extent and directionality is vital to support evaluations and design.

- (Q1152) How can we use these metrics to measure bias and its impact?
- (Q1153) What methods and mechanisms could be developed to determine that an algorithmic process is more (or less) biased than the current state?

Long-term impact is often ignored or difficult to measure and needs to be brought to the surface in order to go beyond the more obvious short-term harms.

- (Q1154) Can we develop a longitudinal study of the impacts of algorithmic decision making on a group's socio-economic opportunities?

Given that human decision making is often based in an individual's perception of ground truth, metrics for comparison against existing processes are needed to support evaluation.

- (Q1155) How are baselines or benchmarks established for current processes to measure potential bias of an algorithm attempting to address that same process?

Mechanisms and evaluative methods are needed to determine whether ML solutions have objectively improved a given process. This visibility could prevent or reduce the use of systems where they are not an improvement.

- (Q1156) What measures and methods would best evaluate how the application of ML improves specific processes?

Exploring the boundaries of bias would help to determine if biases in ML systems are problematic only when dealing with people, or if they would also introduce problematic impacts when dealing with animals, products, environmental concerns, etc.

- (Q1157) How can we determine the boundaries of when bias is problematic and when "statistical discrimination" might not be wrongful?

While bias is an issue in ML systems in general, there are protected classes and dimensions of bias that need the most urgent attention. We need to explore the possible prioritization of bias dimensions and how and if specific statutory determinations can be incorporated to address the priority concerns in each context.

- (Q1158) What are the protected classes or dimensions of bias that need the most urgent attention (e.g., gender, race, religion)?
- (Q1159) Are there unprotected classes that should be prioritized because of the ways in which they might be impacted by systems that might be biased with regard to them?
- (Q1160) Can this research leverage country laws regarding protected characteristics?
- (Q1161) Can we determine if systems (such as facial recognition) can ever be perfectly equally reliable across demographic groups, or if not, what is acceptable or good enough?

Theme E04: Self-reinforcing cycles

There are classes of ML systems that, by their nature, reinforce their own behaviors. Methods research is needed to tie evaluation to ongoing self-reinforcement in ML systems.

- (Q1162) Are there methods that could identify when the possibility of self-reinforcement would mean that the use of a system would be unacceptable?
- (Q1163) When are feedback loops problematic? When are they not?
- (Q1164) When might feedback loops get stuck in a local maximum?
- (Q1165) What measures would quantify the extent to which training data and feedback loops directly impact the very conditions for what is being predicted (e.g., as in predictive policing)?

- (Q1166) How can we measure the impact and rate of change of self-reinforcing systems and provide metrics for their evaluation?
- (Q1167) Would it be possible to identify different levels of severity and harm that can result from self-reinforcing cycles? Are all such cycles harmful or undesirable?
- (Q1168) Which sets of problems and domains are particularly susceptible to the unfavorable outcomes arising from self-reinforcing cycles?
- (Q1169) Is it possible to develop testing simulators aimed at uncovering self-reinforcing cycles?

Theme E05: Outputs, deliverables, and implementation

Going beyond the theory to adoption is crucial to both the evaluation and design of systems. It's necessary to determine the deployment approaches and processes that would be most effective.

- (Q1170) What approaches to evaluation, and presentation of results of evaluation, are appropriate or promising?
- (Q1171) How can we ensure the adoption of a given evaluation scheme if it is proven effective?
- (Q1172) What decisions do we believe should be based on or influenced by "evaluation"? Are they go/no go? "How to"? or something else?

Given the somewhat hidden or subtle nature of language bias, it is crucial to be able to provide an operationalizable solution that can be applied generally.

- (Q1173) Is "risk assessment and mitigation" a viable structure and metric to use to determine evaluation criteria for language bias?

Methods and materials to describe issues of proxy use in training are needed to guide mindful decisions in all three of the functional areas: data, algorithm, interaction.

- (Q1174) What materials would be most useful for training and awareness of the danger of proxies and "discrimination by proxy"?

A clear set of guidelines that exclude ML approaches in fraught areas of application could short circuit the ongoing development of weak and potentially harmful models and focus development on areas likely to significantly improve outcomes and provide positive impact.

- (Q1175) What are the conditions that argue for or against the decision to implement machine learning applications?

Identification of what constitutes trust in various domains provides us with a target for trust as we develop systems.

- (Q1176) How can we provide developers with models of trust and trustworthiness that can form the basis for mechanisms to establish it?
- (Q1177) What areas, sectors or factors are most amenable to mitigation or trust through evaluation and assurance?

Identification of the characteristics of domains and applications that make them amenable to ML would be the first step in the determination of whether a problem is best left to a human, or rather if it is acceptable to delegate it to ML systems.

- (Q1178) What higher-level considerations arch over the use of machine learning in different domains, and which can guide the implementation of machine learning and integration into existing processes?

Much of the human impact of machine learning turns on what happens after an algorithm is developed, trained, and deployed. We need a comprehensive study of how people use and are impacted by different approaches to ML and the systems they support.

- (Q1179) How can we model and analyze what users do with the conclusions, recommendations, or predictions of algorithms?
- (Q1180) How can we train the users who interact with the system to interpret its outcomes with an appropriate mental model of the system?
- (Q1181) How can we develop methods and processes for appropriate utilization and integration of these systems?
- (Q1182) What are the guidelines for the integration of machine learning into the private or public sphere?
- (Q1183) How do we establish expectations for what the technology can do when it is derived from human language, given that human language inevitably contains human biases?

As with other explorations, the identification of bias at the evaluation phase might be used to identify the root causes and to push upstream to refine design approaches.

- (Q1184) Can models be "inverted" to be used to identify the data most responsible for biased outcomes? Might we be able to pinpoint, for example, what subset of the training data is responsible for the correlation between "nurse" and "female," "doctor" and "male" language pairs?

Given the nature of ML in embedded systems, we need approaches to ongoing evaluations in the face of ongoing change.

- (Q1185) How can we develop evaluation mechanisms that are ongoing along a time horizon, rather than treating this process as a one-time checkbox?

Additional Questions: Policy, Regulation, and Education

While identifying research needs, there were several questions that were raised about issues of policy and education, which, while they may not correspond to a specific element of the Framework, could have influence on the human impact of ML systems. While these may be somewhat more open-ended than issues raised in other areas, much of the discussion around the impact of machine learning can end up converging on regulations. As with most new technology areas, discussion around regulation in this space often lacks a clear sense of the length and breadth of the functionality of the technologies under discussion. In order to regulate these technologies, it is crucial that regulators understand them at a functional if not technical level.

The following questions were raised as important to explore further.

- (Q1186) What policies can be developed related to algorithms that can prevent them from being used in harmful or illegal ways?
- (Q1187) How can we define what "abuse" or "inappropriate use" means at an organization/individual level?

- (Q1188) What are the policies we need to mitigate and to correct biases? What strategies, technical and social, do we need and how do we determine them?
- (Q1189) Is it possible to create a public data or model utility that has rigorous standards for data gathering and management and the building and testing of models?
- (Q1190) How can we assure that systems benefit a broad cross-section of the population, not only specific subgroups?
- (Q1191) Can we design organizational incentives to report harms (before or after deployment)?
- (Q1192) Can we establish public disclosure methods, to help solicit input from a diversity of stakeholders?
- (Q1193) Can we develop and enact public education methods, to facilitate and foster informed multi-stakeholder participation?

Appendix B: Roadmap for Research Revision History

Version Number	**Date Issued**	**Summary of Changes**	**Authorship Acknowledgement**
0.0	October 2021	Document origination	This document was compiled by Kristian J. Hammond, Ryan Jenkins, Leilani H. Gilpin, and Sarah Loehr. The content reflects materials and meetings that were held as part of the Machine Learning Impact Initiative in 2020 and 2021, with the participation of a network of researchers and practitioners. See the Machine Learning Impact Initiative Summary Report for a full list of all who participated and engaged in these processes.
1.0	May 2022	• Modified the introduction and executive summary overview text for clarity. • Separated the research questions into a supplementary Appendix A: Roadmap for Research Themes and Questions. • Added additional questions throughout based on discussions since prior version. • Added numbering convention to the themes and questions. • Added Appendix B: Roadmap for Research Revision History.	This version was compiled and modified by Kristian J. Hammond and Sarah (Loehr) Spurlock.
1.1	August 2022	• Added Executive Summary	This Version was compiled and modified by Kristian J. Hammond, Sarah (Loehr) Spurlock, and Thomas Hipchen