C-Accel Track Recommendation: Ethical Design of AIs (EDAIs)
NSF Grant 2232404, PI Louiqa Raschid, University of Maryland
Executive Summary

AIs are pervasively integrated into the fabric of our lives. Disruptive AIs such as autonomous vehicles have captured our imagination and raised concerns. AIs that recommend movies, music or products, prioritize social media posts or search engine results, approve credit applications, diagnose diseases, etc., are rapidly transforming all aspects of our lives. In nearly every domain in which AIs have been deployed, there have been interactions and outcomes that were unexpected and unintended. There is an increasing awareness - across academia, industry, government, and the general public - that active steps need to be taken to ensure that these AI solutions must follow ethical principles that both safeguard, and actively promote, human well-being.

On the bright side, ethics and AI is an active area of research and there is considerable progress in ensuring successful and pragmatic outcomes. Consider the following successes:

- NSF has funded numerous projects on the topic of ethics and AI. (See Appendix)
- There are multiple conferences devoted to the topic, e.g., the ACM Conference on Fairness, Accountability, and Transparency (FAccT); AI, Ethics, & Society; specialized Ethics tracks within broader AI conferences, etc.
- Many companies and federal agencies have hired a Chief Ethics / Ethical AI Officer or a Lead for Responsible AI.
- There is a range of research and workforce training activity across multiple disciplines including ethics and philosophy, the social sciences, anthropology, computer science, etc., as well as in the professional schools for the law, medical and health sciences.
- The US government has convened groups on the topic (e.g., THE NATIONAL AI ADVISORY COMMITTEE (NAIAC) - National Artificial Intelligence Initiative; or a recent National Academies committee on Responsible Computing Research). The US White House has issued a Blueprint for a AI Bill of Rights Blueprint for an AI Bill of Rights | OSTP | The White House.
- Many guidelines and solution approaches have been developed including from the National Academies of Engineering [1], NIST AI Risk Management Framework | NIST and the NIH [54].

However, significant gaps and challenges remain. Large industrial organizations have been able to acquire human capital and strengthen their ethical expertise, but they continue to face the **challenge of retrofitting and re-engineering robust engineering pipelines** that were not designed to address ethical values or satisfy values-based measures and standards. In contrast, small and medium-sized enterprises can potentially design ethical AI systems “from the ground up,” but often lack the necessary human capital. All organizations fear the potential backlash from ethical missteps and would welcome the emergence of a **toolbox of best practices**, tailored to specific domains, and with appropriate measures. More concretely, small businesses, startups, or foundations that support open-source AIs typically do not have adequate resources to build up the needed expertise and staff in-house. They are in dire need of off-the-shelf Ethical AI design solutions, such as a **template AI Governance toolkit and/or checklist** that can be personalized to specific AIs. Outside of the tech industry, Executive Order EO13960 calls for the use of Trustworthy AIs across the US Government; however, there is little consensus about the frameworks, workforce needs, or mechanisms to realize this goal, despite

the proliferation of AI systems across the Federal Government.[1] Even when there is agreement about the desired features of an AI system, the designers, builders, and deployers of AIs often lack the knowledge of how to reach that goal. This is particularly true in areas such as criminal justice, financial transactions, and services, or health and well-being, where there is already significant evidence of human bias, in both training data and outcomes.

There has been much research activity on relevant topics and themes, but more progress is needed, and in particular in the area of **pragmatic and translational guidance**. Although people widely agree that ethical design should focus on stakeholders' values, there has been limited work to translate these values into measures, and then constraints to be applied during design and development, or in understanding how different (heterogeneous, conflicting) values could be explicitly (or implicitly) implemented. Similarly, measures largely focus on "objective" measures such as accuracy or out-of-sample generalization, but the ethical design of AI requires measures that capture the manifold ways that an AI system can support people's values. These two themes are clearly interdependent, as the ideal situation would be to simultaneously understand the nature and measures of stakeholder values. For example, most people value safe driving for autonomous vehicles, but we currently lack clear operationalizations or measures for "safe driving," so we cannot design to ensure that this value is supported. At the same time, the most effective best practices and methods are useless if they are not actually implemented in AI development. We must additionally recognize that organizational (mis)incentives may be a significant limiting factor in the adoption and use of already-known ethical design techniques. Simple exhortations of companies will be insufficient. We need better, more persuasive arguments and frameworks, e.g., demonstrations of positive return on investment.

As governments consider regulatory frameworks, similar to data privacy regulations, independent third-party organizations will be needed to conduct audits and to check for compliance with regulation or certification against standards. Auditors must be equipped to evolve at the same rapid pace as AIs. There may be a need for automated alerting to possible ethical concerns; tactics, techniques, and procedures (TTPs) to allow for rapid auditing; and support for large scale auditing for widely deployed AIs.

A Workshop on the Ethical Design of AIs was convened in September and October 2022; Louiqa Raschid from the University of Maryland was PI, with Michael Pazzani from USC/ISI and John Horty and Ilaria Canavotto from the University of Maryland as co-PIs. https://go.umd.edu/EDAIs

Workshop participants hailed from a wide range of disciplines and application domains, and expressed interest in establishing partnerships across academia, industry, and government agencies, to address the challenges that were identified during the event.

One of the outcomes of the workshop was a recommendation for a 2023 Convergence Accelerator Track on the Ethical Design of AIs (EDAIs). Suggested recommendations of themes and goals for the EDAIs Track include the following:

- Human Centered Design methodologies around Values and Measures and Incentives.
- Proto Ethical AIs: Algorithms or Systems or Pipelines across multiple domains.
- Best Practices for the design of ethical AIs.
- Workforce development and education and training.

This report documents the activities of the EDAIs Workshop as follows: In Section 1, we provide a broad overview of Ethical AI along multiple dimensions. We also expand on the goals for the

[1] In fact, part of the charge to the NAIAC is to provide recommendations about how to achieve trustworthy AI across the US Government. However, those recommendations are unlikely to be at the level of specificity required for particular AI projects.

EDAIs track. Section 2 presents multiple exemplars of use-inspired Ethical AI designs across a range of applications and domains including large language models (LLMs); decision support for health and well being; criminal justice and the management of commercial MLOps platforms. Section 3 provides a research background and a summary of related work. The reports of breakout groups on themes that were explored in the Workshop including Values, Measures, Incentives, and Training and Education and Benefits are included as an Appendix.

1. An Overview of the Ethical Design of AIs

Misunderstandings about the nature of ethical AI pose persistent challenges to its advancement. Three misconceptions are particularly common. First, ethical AI is sometimes thought to be a value-neutral tool. However, the reality is that ethical values are explicitly or implicitly implemented within AI systems as a result of choices made throughout the development lifecycle. For example, the choice of specific success criteria (or loss function for optimization) will result in an AI system that prioritizes some set(s) of values over others. Ethical considerations are therefore already a key part of AI design and development, even when developers are often unaware of the ethical import of those choices. Second, ethical AI is sometimes understood as a matter of compliance, particularly legal compliance. But while some aspects of ethical AI are amenable to compliance certification or other regulatory mechanisms, many others cannot be meaningfully or easily tested through *post hoc* assessments or checklists. For example, many important values, e.g., “drive safely”, “be honest”, and “help others”, cannot be translated into precise performance standards, as their expression depends on complicated details of the exact situation. Further, a focus on *post hoc* assessments may create unintended choices earlier in the design cycle as described by Goodhart’s Law, e.g., a push to optimize the outcome for specific evaluation measures, rather than a focus on a good design that reflects the underlying ethical values. Ethical values must therefore be considered explicitly and critically during the early stages of the design and implementation of AIs. Third, ethical AI is not solely, or even mostly, a matter of abstract philosophical debate and thought. The ethical and societal impacts of AI extend beyond far-future concerns, e.g., superintelligence, or idealized thought experiments, e.g., Trolley Problems. Our workshop focused on the practical, real-world, near-future potential of ethical AIs to lead to tangible improvements in people’s lives. The workshop participants identified numerous projects that could advance these practical challenges and opportunities.

1.1. Ethical AI Principles and Guidelines

Although commonly suggested, a focus on explicit principles was actually deemed by many Workshop participants to *not* be a promising direction for translational research. There have been an enormous number (literally hundreds) of sets of principles that have been proposed to help ensure that our AI systems are ethical (in some sense). There have been so many sets of principles that people have conducted analyses of the similarities and differences between different proposals. The core idea underlying this approach is that clear, careful articulation of the necessary features of an ethical AI system will provide designers, developers, and deployers with a suitable “target” for their work, as well as clear criteria for the subsequent evaluations of their efforts. However, this approach faces four limitations. First, the principles tend to be very high-level, and so usually have limited impact on actual practice. In almost all cases, the descriptions of the principles or desired features fail to translate into practical guidance about how to actually achieve that goal. Second, the high-level nature of these principles means that they must be context- and domain-general, in the sense that they apply in almost all cases. Third, these sets of principles are almost all incompatible with one another, in the sense that there are AI systems that are ethical according to principles A, but not principles

B. This variation in principles thus poses a potential barrier to progress, including the approach of “AI XYZ As A Service”. Fourth, the actual implementation of particular principles has, in practice, almost always been team-sensitive. Different groups, even in the same organization, have interpreted principles in meaningfully different ways.

1.2 Ethical AI Algorithms, Systems, and Pipelines

Near-term progress on AI & Ethics can focus on the development of algorithms (or systems or pipelines) that explicitly encode our values or ethical commitments. For example, responses to Trolley Problem type cases often take the form of algorithms that directly implement one or another theory of ethical evaluation. Additionally, some efforts to respond (ethically) to differing values within a community have explicitly encoded various voting or social preference aggregation procedures in an AI decision system. In some cases, this approach might work well, but it requires two fundamental assumptions that frequently fail to hold. First, this approach assumes that our values can always be precisely and explicitly represented in a machine-interpretable way. In practice, however, many of our values are more vague and context-sensitive. Second, this approach assumes that AI system decisions are made using the same representations and cognitive machinery as humans. However, many AI systems use radically different concepts than humans, so our ethical theories cannot necessarily be translated into AI algorithms. For example, most self-driving cars do not represent the age of pedestrians, or even that there are pedestrians, as opposed to “volumes of space that should not be entered”, so many of the proposed ethical algorithms to solve “Trolley Problem” style cases simply cannot be implemented in the AI systems. More generally, AI systems often find patterns in our environments that we have missed, but those patterns will typically require the AI system to think differently than we do, and thus our ethical theories cannot be directly implemented in such systems. Workshop participants discussed the challenges, needs, and gaps of use-inspired and domain- or application-specific proto-AIs.

1.3 Verifiable Behavior

One could instead focus solely on the behavior of the AI system: does it act in ethical ways (regardless of exactly how those behaviors are generated)? One manifestation of this approach has been the development of test, evaluation, and audit frameworks. For example, the Department of Defense (including multiple branches) has worked to develop test, evaluation, validation, & verification (TEVV) procedures for their AI systems, including those procured from third-party vendors. Many efforts around algorithmic audits, particularly for biases, similarly focus on system behaviors (decisions, classifications, predictions, etc.) rather than the mechanisms by which those are generated (e.g., NYC Local Law 144 on automated employment decision tools). A different manifestation of this approach focuses on systems that are provably beneficial or reliable, including efforts to translate frameworks such as zero trust security into AI contexts. This approach typically requires a high degree of specificity–perhaps implausibly high–about what constitutes “ethical behavior.” In practice, we often do not know exactly which behaviors are most ethical; at the very least, there are typically blurry lines between ethically obligatory and ethically permissible behaviors, and between ethically permissible and forbidden (i.e., unethical) behaviors. While there have been some successful efforts to specify ethical behavior, these almost all arise in relatively closed-world systems, i.e., where most or all relevant factors can be represented, though not necessarily measured.

While the issues around formally verifiable ethical behavior and outcomes formed a backdrop for many of the discussions, the EDAIs Workshop did not focus on formal verification since a companion workshop on Provably Safe and Beneficial AIs (PSBAIs) explored these topics in significant detail. https://humancompatible.ai/psbai-workshop-2022/

1.4 Human-Centered Ethical Design and Best Practices

There is a need to develop “best practices” for each stage of the human-centered Ethical AI lifecycle – design, development, evaluation, deployment, revision, etc. – that increase the likelihood that ethical AI will result. Importantly, these best practices do not each need to explicitly consider ethical issues, i.e., they should be judged by whether they lead to more ethical AIs, not whether they use “ethical” language. For example, the best practice of “engage with diverse communities to determine how an AI system might affect them” is difficult to express in purely ethical language or to capture as an ethical value. Nonetheless, this best practice has been shown to consistently lead to more ethical AI systems. Best practices have been shown to have positive impacts in many domains, and best practices can usually be adopted and deployed in industry contexts. Discussion at the workshop centered on identifying challenges and opportunities for the development, dissemination, and widespread adoption of these best practices.

1.5 Potential Themes and Deliverables of a Convergence Accelerator Track

We here describe four high-level clusters of projects, challenges, and opportunities that emerged from the Workshop.

- Human Centered Design methodologies around Values and Measures and Incentives.
- Proto Ethical AIs: Algorithms or Systems or Pipelines across multiple domains.
- Best Practices for the design of ethical AIs:
    - Toolkits for designing, implementing and assessing AI systems.
    - Platforms for AI Governance.
    - Independent third-party audits for compliance and certification.
- Workforce development and education and training.

Human-Centered Design Methodologies:

The EDAIs Workshop articulated the need for human-centered and pragmatic design methodologies around the themes of Values, Measures, and Incentives. Workshop participants explored high-level guiding principles, while simultaneously identifying projects to ground them in best practices. They identified the key challenges within each theme, and the relevant research and design questions to address the challenges. They went on to identify the needs, gaps and the obstacles that must be satisfied or overcome, to be successful. The detailed outcomes of the discussions are in an Appendix.

Ethical AI systems must support the values of key stakeholders, but there are few concrete methods, tools, or frameworks for systematic value elicitation, or for the translation of values into measures, or for the generation of constraints during design, development, and deployment. For example, simple prototypes are useful for helping non-technical individuals to identify and articulate their values, but there are few systems to port values into measures that can apply to real-world prototypes. The ethical design of AI systems would be greatly enhanced by tools and processes that can translate natural language expressions of values, i.e., what people naturally produce, into design- and development- stage measures and / or constraints. Deliverables of an EDAIs Convergence Accelerator Track could range from validation of relevant methodologies, to the development of tools and APIs, to use cases and training material.

Proto Ethical AIs:

EDAIs Workshop participants were invited to explore use-inspired and domain- or application-specific Proto Ethical AIs. Participants addressed the following questions: What are the major ethical and social concerns for this specific application domain? How can ethical

design address these concerns? What major gaps are left unaddressed? Several of these use cases are presented in Section 2.

A Convergence Accelerator Track could include Proto AIs that range from compact modular systems to complex engineering pipelines. Proto Ethical AIs can help to uncover the potential mismatch(es) between the representations and cognitive machinery employed by humans versus those implemented within the AIs. Deliverables could include an examination of how well AIs can (or cannot) articulate values, implement measures, provide assessments, and reflect incentives that promote human well-being.

Best Practices - Toolkits for the Design of Ethical AIs:

There was much discussion and enthusiasm around the development of Best Practices and Toolkits for the design and assessment of ethical AIs. This is a fertile opportunity for partnerships across multiple stakeholders. Deliverables could include the following:

- The development of best practices and protocols around datasheets, model cards, triage checklists, etc. [55,56].
- Best practices for the design of systems that mitigate ethical issues. Plenary speakers Kearns and Etzioni described examples of such systems, e.g., systems that learn accurately for protected minority classes [57] and systems that exhibit common sense reasoning about ethics [58].
- Tools for testing data for representativeness. Tools that assess systems for ethical issues, e.g., a higher error rate in protected minority classes.

Best Practices - Platforms and Support for AI Governance:

There are close connections between design and governance. Effective design requires some understanding of the goals that governance helps to achieve. Governance must be sensitive to opportunities and more important, constraints, during design and development. AI Governance can also vary dramatically based on the size and complexity of the AIs as well as the organizations that are developing or deploying these AIs. The following deliverables of a Convergence Accelerator Track would help to guide or strengthen AI governance:

- Systems to identify relevant regulation. Frameworks or approaches to comply with regulatory requirements, e.g., appropriate methods to audit for biases.
- Protocols and checklists for Governance-in-a-box solutions that can be readily deployed. Approaches for customization of in-a-box solutions to specific application domains.
- Governance deliverables could range from simple checklists to APIs and services to powerful sandboxes for training, testing and evaluation.

Best Practices - Independent Third Party Audits for Compliance and Certification:

The government has an important role to play on behalf of consumers, in particular for applications and domains such as medical diagnostics, credit scoring, sentencing or parole decisions, etc. They can do this by establishing standards or through regulatory frameworks. Independent third party audits are then needed to ensure that AIs are compliant with regulations and/or meet standards and certifications. Deliverables could include the following:

- Tactics, techniques, and procedures (TTPs) to allow for rapid auditing.
- Automatic measures that can evolve at the pace of new product updates and releases.
- Support for large-scale auditing, and customization.

Workforce Development, Education, and Training

Training and education are needed to develop interdisciplinary collaboration skills. There is also a need for workforce development, e.g., training for roles within ethical AI ecosystems. Workshop participants highlighted some of the following objectives and deliverables:

- Educating the public about the benefits and the potentially harmful limitations of AIs.
- Leveraging lessons learned from other domains or historical inventions, to build an understanding of current AI technology & best communication practices.
- Training technologists in ethics and training ethicists about technology.
- Providing the relevant training for non-technologists who are professionals in the law, regulation, and compliance, or in domains in which AIs are extensively deployed, so they can contribute meaningfully to ensuring positive outcomes and minimizing harm.
- Educating human-centered designers and users on the need to address the landscape after the successful deployment of an AI.
- Deliverables in this area could include the following:
    - Repositories for courses and curriculum.
    - Data sets that illustrate problems/challenges in the design of ethical AIs.
    - Case Studies on ethical failures, how they occurred, and how they can be prevented or mitigated.

1.6 The Ethical AI Ecosystem

Workshop participants hailed from a wide range of reference disciplines and application domains, and expressed interest in establishing partnerships across academia, industry and government agencies, to address the challenges that were identified during the event. Below is a list, undoubtedly not exhaustive, of groups and individuals who have expressed interest in something like an EDAIs Convergence Accelerator track.

Academic Researchers:

- Computer Science and Engineering: An interest in AI systems development or research in AI technologies or human-centered design.
- Social Sciences: Apply theories and methods from the social sciences to study the impact of technology on users, organizations and society.
- Humanities, including Philosophy: Explore foundational issues about values, ethics, and the broader impacts of ethical AIs.
- Human-Computer Interaction and design.
- Law and Public Policy: Legal scholarship, regulations, and compliance.
- Professional disciplines - Medicine, Public Health, Business, and Management, etc.: Domain-specific expertise in both AI technology as well as AI ethics.

Industry:

- Information technology companies that design and deploy AIs (Google, Meta, Microsoft, Amazon, etc.) or companies that design and manufacture tools and devices and instruments (Intel).
- Companies that are involved in the deployment of AIs in specific domains including finance (FICO, Mastercard, Experian, JP Morgan Chase), entertainment (Netflix, Hulu), legal services (Thomson Reuters), defense (Boeing, Lockheed Martin), healthcare (GE Health), consulting and auditing (big four US consulting firms).
- While large public companies get the most visibility, there are many successful small

businesses and startups in the field (HuggingFace, Humanyze, Distributed AI Research Institute, AI Ethics Lab, Redgrave Data).
- Developer PaaS companies, e.g., AWS, Azure, Google Engine.

Government can play a range of roles in the design and use of Ethical AIs:

- As a procurer and consumer of Ethical AI technologies.
- As a source of training and ground truth data.
- As a developer and enforcer of regulations as well as best practices.
- As legislators creating laws that enforce AI ethics, e.g., data privacy such as GDPR.

Participants from the Departments of Justice and Defense played a key role in the EDAIs Workshop and were interested in the above activities.

- A convener of researchers in the Ethical AI space, producing reports and recommendations, e.g., OSTP and NIST.
- Agencies that fund Ethical AI include scientific agencies (NSF, NIH, NIST, DARPA, IARPA), government research laboratories such as the Naval Research Laboratory, and mission-focused agencies including the Department of Defense, Agriculture, and the Department of Education.

Not-for-profit companies and foundations:

- AI Labs such as the Allen Institute For Artificial Intelligence, Open AI, and Machine Intelligence Research Institute.
- Foundations that fund research on Ethics and Society such as the Mellon Foundation, Open Philanthropy, Schmidt Futures, Ford Foundation, MacArthur Foundation, and Omidyar Network.
- Advocacy groups such as ACLU, Leadership Conference for Civil and Human Rights, Color of Change, Movement Alliance Project.
- Technology policy organizations such as the Center for Democracy and Technology, Upturn, Algorithmic Justice League.
- Nonprofit research institutes such as the Data & Society Research Institute, and Ada Lovelace Institute.

Independent third-party AI auditors and reviewers:

- Associations and companies that provide consumer protection, independent review and testing and safety, such as the Underwriters Laboratories, Consumer Reports, yelp, etc.
- Technical reporters such as the Vox, Wirecutter, the AI Incidents Database, etc., and review sites such as yelp.

International Organizations:

- OECD, UNESCO, WEF, GPAI, PAI, CERN
- EU, Council of Europe, UN

2 Use Inspired Ethical AI Design Exemplars

2.1 Large Language Models

Large Language Models (LLMs) such as GPT-3 [43] are trained on very large databases of text usually found on the internet, such as Wikipedia or social media data. They tune millions to billions of parameters that allow a model to predict what words might be relevant in a certain context, given a prompt. They can be used for a variety of purposes including machine translation of text, correcting grammar, auto-completing words or sentences, question answering, generating new bodies of text such as news articles and essays, and beyond [Brown]. The set of potential applications has grown rapidly in recent years to encompass attempts to write scientific papers (e.g., Galactica), to engage in creative word-based games (e.g., AI Dungeon), to act as virtual teaching assistants (e.g., Jill Watson), and for code generation (e.g., Codex).

https://www.technologyreview.com/2022/11/22/1063618/trust-large-language-models-at-your-own-peril/

However, LLMs can also be used for malicious uses, such as the mass generation of false news and misinformation.

https://cset.georgetown.edu/event/large-language-models-and-the-future-of-disinformation/

The use of LLMs through chatbots has been particularly controversial, demonstrating certain key ethical problems, as LLMs can reflect whatever biases are in the corpus they are trained on. Further, with billions of parameters, it is difficult to identify and correct for numerous types of biases related to gender, race/ethnicity, language, cultural appropriateness, etc. The following examples illustrate such problems:

- Microsoft's AI Twitter bot Tay was pulled after posting racist and sexist tweets [44]. In this case, the bot learned to incorporate and react to people's conversations on the internet without considering whether modeling these internet users was appropriate. Numerous studies have found biases in LLMs [45].
- Further illustrating the potential harms to rights and safety, a Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves [46]. The problem exists because although statistically generating plausible replies may appear to be reasonable on the surface, LLMs lack the knowledge, common sense, or human experience to understand the implications of the text that they generate.
- After the workshop concluded, yet another large language model (Galactica) was released and shut down in two days because "it spewed misinformation." https://www.cnet.com/science/meta-trained-an-ai-on-48-million-science-papers-it-was-shut-down-after-two-days/

Additional concerns involve the representativeness of data, transparency of the models, and human accountability as third-party vendors and open source LLMs are increasingly used in the public and private sectors.

What are the Ethical and Societal Concerns?

- What is the appropriate way to delineate the responsibilities of LLM providers versus users?
- How can we open up access to powerful models for beneficial purposes, transparency, and accountability, while minimizing the potential for malicious uses like misinformation? What are the appropriate guardrails, conditions, or legal contracts to balance these

issues?
- With respect to the ethical risks of LLMs and understanding them, such as biases, what constitutes sufficient testing? How do we measure domain-specific performance (e.g., LLMs in a legal setting versus in a healthcare setting)? How do we measure uncertainty in the quality of LLM output?
- What are the possible emergent capabilities of LLMs? How can we make them more predictable or controllable?
- How can interdisciplinary teams or approaches be used to evaluate the risks and implications of LLMs? What is the role of interdisciplinary thinking here?
- When should humans be in the loop regarding LLMs? During the training process, during use of LLMs?
- What kinds of data are appropriate to use in LLM training? What are best practices in data acquisition and cleaning, including considerations about the ethics of sourcing public or copyrighted data?
- What types of filters can be applied to ensure LLM output is appropriate, e.g., private information filters, cultural sensitivity filters, filters about race/ethnicity or gender representation, etc.?
- What types of normative constraints, human common sense, or human-in-the-loop training processes can be leveraged to improve LLM quality and ethical soundness?
- How should LLM developers collaborate to help collectively manage LLM development and use, in light of competition and trade secrecy concerns?
- How can LLMs be improved to serve different language speakers, or operate effectively and appropriately in different regions and cultures?
- How do you track the carbon footprint in training or fine-tuning LLMs, especially when LLMs are trained via cloud service providers? What would a balanced framework look like that does not cause excessive environmental harm while also preserving the capacity for experimentation and invention? Relatedly, what are the article costs for post-deployment model monitoring or explainability techniques used outside of the initial model training process? And how often is it appropriate to re-train models?

How can Ethical Design Help?

- Creating better tools and accepted pipelines/workflows for LLM development could improve the sustainability, replicability, traceability, and trust of LLMs
- Providing open source infrastructure could facilitate increased research by academics and civil society, and allow for testing and transparency
- Identifying a “CI/CD” or a continuous way to introduce ethical testing and make available, e.g., as a standard across industry and academia, could allow us to test violations against ethical principles
- Promoting open source methods for responsible LLM development (e.g., red teaming, test kitchens) could allow for more collective oversight and learning

What Gaps Remain?

- AI ethics research should expand its focus beyond data science, to include AI pipelines and software engineering approaches. This will help to understand key issues and identify relevant solutions.
- Powerful LLMs may increasingly constitute mission-critical systems, meriting heightened rigor for their safety, efficacy, and impact.
- Conversations on ethics and safety (and across social science, humanities, CS, and engineering) can be united to provide a more holistic understanding of LLMs.

2.2 Criminal Justice

Decision making in the criminal justice domain raises significant issues of ethical and societal concern. While the deployment of AIs can further propagate existing human biases and potential unethical outcomes, there is also the potential that the introduction of AIs can mitigate or overcome some of these challenges.

Consider the very common scenario of police decisions to search an automobile for drugs or to detain the auto for a canine search. The ability to use NLP/ML approaches to analyze a large number of cases may result in the identification of factors on which courts and police may reasonably rely, to determine whether an officer has reasonable suspicion to search or detain the automobile. Eventually, such findings could assist federal, state, and local jurisdictions to promote a fairer administration of drug laws in the criminal justice system. It could also provide AI tools to guide future police and judicial decision-making. One benefit would be in collecting data with which to assess whether the police/judicial decisions and the factors they rely upon actually correlate to the discovery of drugs in a vehicle.

The design of such tools for guiding police decision-making in the field raises potential ethical design issues to be discussed.

What are the Ethical and Societal Concerns?

- How can one anticipate and identify the potential for misuse of the tool such as gaming the system by police officers?
- How to track and manage how user inputs may be biased? This is an acute issue in a domain that has given rise to the phenomenon of “driving while black”.
- How to teach users about the limitations of the AI system and its proper use? Training police in how to use the tool, how it can help them, and why it is important to use it properly will be a key challenge.
- How to deal with users’ over or under trust of the system’s recommendations?
- How to avoid the effects of presenting predictions with respect to priming users or biasing their judgments?
- How to track conditions that require a change of control between the system and a user?

How can Ethical Design Help?

- A key is to understand the contexts in which users (i.e., police officers or judges) would be using these tools. What problems are they dealing with and how could the tools help them, while at the same time encouraging accurate inputs and recording truthful data?
- For example, police also may be recorded orally and visually in a way that provides a kind of reality check against which their data entries could be assessed for accuracy.
- To better understand the factors and how police and courts apply them, it would be desirable to assess the case texts for racial or ethnic bias (explicit or implicit) to see their effect on decision-making.

What Gaps Remain?

- Identification of explicit or implicit racial bias in case decisions should be technically feasible, but it is an empirical question.
- Enabling NLP/ML to identify factors in case texts is feasible but can it be done well enough to enable reliable statistical analysis of factor weights?
- Assuming these technical challenges can be met, can the results inform policy decision-making, enable building a tool, and support acceptance of such a tool by police departments and the judiciary?

- It will be important to engage the law enforcement community early in the design process and to identify whether the above ethical concerns/challenges can be met.

What are the incentives to build an ethical AI?

- The legal standard governing police/judicial determinations of “reasonable suspicion” is vague. There are too many cases for judges or police to read or take into account.
- If we can compute the weights of factors from many cases, we could bring a level of objectivity to these determinations and use it to provide guidance through a tool. The data the tool provides about factors and decisions could then be related to whether drugs were actually found. This could lead to an objective policy assessment about whether the factors and decisions make sense.

2.3 Health and Well-Being Decision Support

Artificial intelligence has the potential for improved healthcare outcomes. AIs can assist doctors during diagnosis or in specialized tasks such as image analysis. Genomics can also lead to personalized medical care [47]. Currently, most medical AI systems make recommendations to clinicians who can use their judgment to override, but not to collaboratively come to a conclusion. Medical AI systems have the potential to reinforce and amplify existing biases [48, 49] and perhaps to introduce new and difficult to detect biases [50]. Consider a well-known example where machine learning reinforced existing biases against African Americans in healthcare [51]. An analysis of the cause of this bias found that an initial system was designed to reduce the overall cost of healthcare rather than maximize the outcome of patient health. As a consequence of this design choice, some patient populations were underserved. They were not recommended for (more expensive) treatments that may have led to better outcomes; this consequently led to lower healthcare expenditure for these patients. Unfortunately, the context of cost reduction as a goal was not considered when learning from this patient population. The AI learned that this patient population did not need as many treatments and their healthcare costs were lower. The NIH has begun to incorporate some ethical principles in its research community, e.g., by requiring that research data is findable, accessible, interoperable, and reusable (FAIR). However, to reap the benefits of AI, without the possible negative effects, designers need to take transparency, fairness, representativeness, and explainability into account, at all stages of the treatment pipeline.

What are the Ethical and Societal Concerns?

- A need to be explicitly inclusive in creating datasets for AI algorithms. The need for training of computer scientists on the impacts of choices made at early stages such as the curation of training data.
- Getting representative data (age, race, gender, ethnicity). (Melanoma database example). Data collected in low-income communities is often not as well-integrated into electronic health record (EHR) systems that are the major source for training data. Representative sampling is a huge challenge and this is a domain where convenience samples are extremely problematic.
- Medical AI systems often use sensitive information without explicit notification or understanding of the impacts, e.g. identifying gender or race from retina images may lead to non-representative data.
- Decision-making on mental health issues can be judgmental. How do you define “harm” and when to trust someone’s judgment, e.g., in suicidal contexts?
- Misdiagnosis and its impacts. An example is an awareness of the complexity of

diagnosis, e.g. along the autism spectrum. Differences in the racialization of disease, e.g., African American children are less likely to be diagnosed. Cultural differences in finding specialists.

- Health Hazards introduced by AI capabilities (AI functionality controlling safety-critical health data, statistics, and making decisions on the health of people)
- How Data (integrity/assurance) is captured, analyzed and verified/validated to make health decisions? (Is/can data be defined as “Safety Significant (critical) data”?
- Related to rare data, images, etc. can synthetic data and different approaches to evaluation help?
- AI is more likely to learn existing practices and biases, instead of discovering and correcting for such biases.
- ML methods need to be carefully employed in domains where causal reasoning is important (e.g. when predictions are based on past data generated by unknown policies which could be biased, simple, etc.)
- How do we define and validate the measurements that go into all the AI algorithms (labels, use of predictions, etc)?
- There are some serious concerns with the undesirable outcomes of well-intended decisions by humans and the use of such outcome data for labels. This is actually related to defining ‘harm’ versus ‘no-harm’ or ‘hate speech’ versus’ non-hate-speech’
- How do you define “do no harm” for training or coding AIs Including physical and other harm?
- Concerns about agency in AI-Brain interfaces. Data privacy. Who is reviewing and testing the tech in brain/AI interfaces, and under what governance model (FDA not knowing how to deal with AI)

How can Ethical Design Help?

- Standardization of policy. Standards addressing Ethical Principles across all disciplines would be a good first step to making progress in this domain.
- More qualitative and behavioral research to identify the potential for harm during disease diagnosis and disease management, e.g., a lack of diversity and the resulting undesirable harm across mental health cases, including autism, dementia, age-related dementia, etc.
- Develop AI methods that can handle different levels of uncertainty across labels.
- Clear specification of the explicit uses of AI capabilities as well as ensuring that there is appropriate testing and assurance within those defined uses across the entire life-cycle.
- Establish the expectation and standards for ethical review and testing of AI systems before deployment.
- Establish standards (thresholds, cases, applications) to differentiate when anAI system could potentially substitute for human judgment versus when an AI system should only augment/assist in human decision-making.

What Gaps Remain?

- Gaps between data capture and analysis methodology, and gaps when interpreting results to make decisions on safety and risk.
- Data for mental health and neurological diseases are still not sufficient.
- Need more training data and methodologies for student causal-effect analysis.
- Raising awareness about the harm caused to target subpopulations, in particular, protected minority groups.

2.4 Increasing Ethics in Commercial MLOps Platforms

Because of the promise and advanced capabilities of AI/ML (AI models trained using ML), many start-ups aim to train and integrate AI into their end-user applications. Often, businesses procure MLOps platforms to accelerate their data scientists' ability to quickly train, test, and then deploy AI models into operations. However, the MLOps platforms currently provide basic features and can be improved to help foster more ethical and socially responsible AI.

We focus on ethics and data privacy. Today MLOps platforms will train models using whatever data is provided – regardless of whether these data contain private or sensitive data (e.g., faces with associated names) which would violate new data privacy regulations. In addition, there is no means in the MLOps platforms to trace the provenance of personal data within training data (e.g., photographs of individuals, text written by certain authors) so that these data can be removed from the model upon the person's/author's request. There are no means for individuals or creators to opt out of having their information used for specific AI model development.

Additionally, current MLOps platforms do not automatically create domain testing criteria nor do they test that the optimization criteria considers the tradeoffs of benefits to individuals, groups, and societies. In addition, MLOps platforms do not integrate well with developers' DevOps platforms to ensure that AI models are understood and correctly used by end-users through the application interfaces.

What are the Ethical and Societal Concerns?

- Since MLOps platforms are used to create AI models from data, how are these platforms helping to ensure AI ethical issues for individuals and societies?
- How can one obtain explicit permission to use one's personal information for training an AI/ML model?
- How can one remove data from the dataset and have the AI model forget what it has learned from the data – without retraining from scratch?
- How can we obtain a consensus about what models are ethical to build and which should not be attempted?
- How does one measure ethics and societal concerns so that they can be integrated into the optimization criteria of AI/ML?
- How can one create optimization criteria that trade-off the competing goals an ethicist must consider – including balancing individual freedoms with societal benefits?
- Who is responsible for deciding on how these tradeoffs should be weighed?

How can Ethical Design Help?

- Having individuals knowingly opt-in (i.e., giving permission to use their data) to train specific AI models will provide evidence of compliance with data privacy regulations
- Having individuals knowingly opt-in can also be used as evidence for the ethical acceptance of the AI model's purpose
- By creating measures of MLOps ethics, companies can have an independent assurance that they are utilizing best practices in their AI/ML design and implementation

What Gaps Remain?

- Explicitly tracking the training data provenance and permissions for any personal information.
- Development of mechanisms within platforms that deny usage of data that have no verified opt-in metadata associated with them.

- In those instances where data provenance is unavailable, the ability to test models to uncover the inclusion of unallowable data that needs to be removed.
- Being able to remove training data from the dataset upon request of individuals, ensuring that the AI has forgotten this data without retraining the AI models from scratch.
- Optimization criteria that consider ethical tradeoffs between individual, organizational, and societal concerns – and who has the responsibility to make these tradeoffs.
- Considering the end-to-end development, there are three primary areas where there are substantial gaps that need to be filled regarding ethics in AI.
- Issues related to how data are gathered, cleaned, normalized, and harmonized against the task at hand and desired learning outcomes.
- Issues related to how algorithms are selected, specific features of the data are selected, and the model is iteratively trained and tested.
- How are user interactions designed and developed to facilitate functionality and usability? While post-ML training, there are questions that developers need to be able to answer as they impact the performance of the systems into which the models are embedded.
- How can we remove biases and data privacy issues from open-source data sets and AI models that serve as foundational models used on MLOps platforms?

What are the incentives to build an ethical MLOps platform?

- Because of societal pressures, many companies would like to know that they are doing everything possible to create ethical and equitable AI – to build trust with their customers.
- New regulations on data privacy will push industry to analyze its collection and use of data
- MLOps platforms can provide independent assessments that ethical AI development of best practices are being observed
- Independent testing and certification of MLOps platforms could assure the platform customers that using the platforms will help their data scientists create AI/ML faster and more ethically – in a manner that garners their customer trust.

3. Research Background on Human-Centered AI, Ethics, and the Law

Researchers in AI and related communities, both academic and industry-focused, have become increasingly concerned with the social and ethical dimensions of AI. This concern has led to a new field of investigation, labeled as Ethics and AI, sometimes as Human-Centered, Human-Compatible, or Humane AI. Regardless of labeling, the goal of this field is to develop the conceptual and technical frameworks that are needed to advance AI in a way that is not only ethical but also promotes human well-being.

Multiple overlapping groups of issues and approaches have emerged in recent years. For simplicity, we summarize them in the following sections, without attempting to provide an integrated or unified roadmap.

- Over the last decade, a range of organizations have published guidelines or policy frameworks to stimulate progress in the application of ethical and social principles. We summarize these guidelines and their limitations.
- There has been significant activity in the AI and ML community around the development of AI systems that are trustworthy. Key characteristics of trustworthy AI systems include transparency, fairness, representativeness, explainability, algorithmic accountability, human control, and privacy.
- A major goal of Human Centered AI centers around the design of autonomous systems capable of reasoning with laws, regulations, and ethical norms. This challenge spans the fields of knowledge representation (KR) and machine learning (ML) in computer science but also involves central issues in moral and legal philosophy. We summarize the main approaches for the design of AI systems that can acquire, represent, and act on the basis of normative information, such as ethical principles or social and legal norms.
- An important set of issues involves the development of appropriate legal and regulatory frameworks for the development of AI systems, as well as techniques for verifying that the resulting laws and regulations are satisfied.
- A final section deals with specific Design Guidelines. This includes strategies that incorporate human-centered design principles into the design of User Interfaces (UIs). Another set of guidelines address the tasks of defining objective functions, datasets and metrics, to test the success and limitations of prototype AI solutions in some selected domains.

We note that we refer to the concept of improving human well-being as it applies to the population at large. We recognize that different population segments may have very different experiences with respect to ethical and social norms. This includes young adults, under-represented or marginalized communities, as well as segments that have been disproportionately impacted in a negative manner with respect to access to credit, or health outcomes, or incarceration or exposure to violence. It has been well recognized that the introduction of AIs can further have a negative outcome on these segments [2, pp. 34-39]. This issue is addressed briefly when we consider evaluation metrics, training data, etc. but it is not explored in depth. We expect to address this more fully during the workshop but we note that this is an important topic that merits a separate line of research and best practices.

3.1 Ethical Guidelines: Benefits and Limitations

Over the last decade, international and national organizations ranging from the Association of Computing Machinery (ACM) to the US National Academies of Science [1] to the High Level Expert Group on Artificial Intelligence appointed by the European Commission to the expert group on AI in Society of the Organisation for Economic Co-operation and Development (OECD) to the National Institutes of Science and Technology (NIST) [3] have published guidelines or policy frameworks [4, 5].

Eleven overarching ethical values and principles have emerged from the content analysis of

over eighty documents reported in [4]. These are, in decreasing order of frequency of the number of sources in which they were featured, as follows: transparency; justice and fairness; non-maleficence; responsibility; privacy; beneficence; freedom and autonomy; trust; dignity; sustainability; and solidarity. The first five, from transparency to privacy, are the most discussed values and principles. We note that transparency and fairness, the two most frequent values, are also the values most studied within the AI and ML technical communities. However, values such as representativeness, explainability and algorithmic accountability did not attract much attention in the policy guidelines.

While relevant and comprehensive, the majority of these guidelines have two serious shortcomings that limit their impact. The first is that the guidelines are typically identified in very abstract terms, with high level goals, but lacking in the granular details that may lead to specific design criteria or other types of constraints [6]. Adding to this shortcoming is that these principles are each explored independently, with little effort made to understand when they may lead to potential conflicting scenarios, in specific domains or AI solutions. There is also no discussion of any path forward to address conflicts, e.g., providing priority to one principle over another for specific use case scenarios.

3.2 Trustworthy AI

An important set of challenges involves the ethical and social problems presented by the increasing sophistication of AI and its prevalence in society. Many of these problems are amenable to traditional avenues of investigation from ethics and the social sciences. However, they have also led to important new areas of technical investigation within computer science itself, especially on how principles of trustworthy AI can be incorporated into the design phase and formally represented so that they may be verified. In particular, enormous technical efforts have been undertaken in machine learning to meet ethical targets like transparency, fairness, representativeness, explainability, algorithmic accountability, promotion of human agency, and preservation of privacy:

- Transparency: Indicates the availability, to a given stakeholder, of sufficient information about how an AI system works and more importantly, if it can be reproduced [7, 8].
- Fairness: Refers to the goal of minimizing algorithmic harms deriving, for instance, from algorithmic bias [9, 10].
- Representativeness: Refers to the extent to which the data used to evaluate (and in some cases train) an AI system matches the situation in which the system will be deployed [11, 12].
- Explainability: Refers to the problem of making it possible for human users to understand and justify the output created by machine learning algorithms [13, 14].
- Algorithmic accountability: Refers to the problem of ascribing responsibility for discriminatory and inequitable outcomes caused by AI systems [15].
- Human agency: Refers to the problems of determining how much control human users should have over AI systems and of developing AI methods that give humans the appropriate amount of control [16].
- Privacy: Refers to the problem of developing regulations, practices, and technical features of AI systems that aim at protecting the privacy of individuals given that ML systems often process personal data.

3.3 Knowledge Representation and Reasoning

There are two general approaches to the problem of designing autonomous systems capable of reasoning with laws, regulations, and ethical norms, with, again, well-known advantages and disadvantages. The first is the top-down approach, according to which normative information is explicitly encoded in a symbolic formalism, such as a logic programming language [17–20] or a deontic logic [21, 22]. The main advantage of this top-down approach is that the symbolic

representations it relies on tend to support a style of computation that leads to transparent, explainable decisions. The central disadvantage of the approach is that it is simply not realistic to imagine that any significant body of normative information could be encoded by hand, due to the exception-laden nature of normative rules and the fact that these rules are often stated using open-textured predicates, which would require further interpretation.

Standing in contrast to the top-down approach is the bottom-up approach, according to which, in its more usual formulations, normative information is acquired through ML techniques, such as reinforcement learning or inverse reinforcement learning [23–25] and encoded, for example, in a reward function or in a distribution of weights in a neural network. The central advantage of this bottom-up approach is that it avoids the knowledge acquisition bottleneck—complex normative information need not be hand-coded but can be extracted from the training data. Further, the ML techniques at work in typical bottom-up systems have proved to be strikingly successful in other domains, such as pattern recognition, facial recognition, and text understanding. It is therefore not unreasonable to hope that these techniques might allow a machine to learn complex moral information as well.

The central disadvantage of the bottom-up approach is that, although learning may indeed take place, it is often unclear exactly what normative information has been learned: how are decisions based on this information supposed to be explained or justified?

Because of the difficulties facing pure top-down or pure bottom-up approaches to the acquisition and representation of normative information, a number of researchers have begun to explore hybrid approaches, combining explicit symbolic representation with machine learning. These hybrid approaches have been developed in different domains and adapted for different reasoning tasks. For example, one early, well-known system, initially explored in the bioethical domain, but then extended to several others [26, 27], represents particular decisions as vectors, with the vector components standing for the extent to which various prima facie moral principles are satisfied or violated, as a result of that decision; these decisions are classified as right or wrong by domain experts, and then the general rules thought to guide this classification arrived at through inductive logic programming. More recently, it has been suggested [28] that a particular hybrid architecture might help medical professionals make allocation decisions for organ donations. On this approach, morally relevant features of potential donor recipients are first identified by domain experts; preferences over competing clusters of these features are elicited from members of a population, and on the basis of these preferences, ML techniques allow the system to offer recommendations.

3.4 Compliance and Verification

In the fields of Human-Centered AI and, more generally, human-robot interaction, four main approaches to the problem of validation and verification have emerged, with well-known advantages and disadvantages:

- Formal verification: Relies on either theorem provers [29–31] or model checkers [22] to exhaustively examine all of a system's possible choices. There is an underlying limitation that this is appropriate when one can provide a largely simplified representation of the environment.
- Simulation-based testing of human-computer interactions [32] can be carried out against the background of a more realistic environmental model. A limitation is that it only simulates human-computer interactions rather than considering real ones.
- User evaluations [26, 27] are based on the evaluation of real observations of human-computer interactions. They have been controversial since users are often deeply divided on fundamental issues of moral import as well as meta-ethical intuitions.
- In light of this, a number of researchers have begun to explore an approach that

combines different verification and validation techniques, to tackle the analysis of safety in human-computer interactions in a more holistic manner [33].

3.5 Human-Centered Design Principles

A strategy to design AI technologies that centers around human capabilities and involvement is presented in [34]. It starts with a change of design metaphors, e.g., from intelligent agents to AI-infused tools, or from social robots to active appliances and moved on to AI operation and control centers, in the spirit of Network Operations Centers. We focus here on the design of user interface (UI) guidelines for ensuring human control and human-centered objective functions, datasets and metrics.

3.5.1 User Interface (UI) Guidelines: UI guidelines should provide users of AI-infused tools and active appliances a greater understanding of the state of the machine, its step by step behavior, and its potential for failure. Users require feedback (e.g., via inclusive visual, auditory, or haptic previews) so they can control execution, similar to the control of cameras or navigation systems. The UI guidelines must accommodate a range of users and expertise; some may wish to have a simpler interface with less options and controls, while others may desire a greater level of feedback granularity and greater control over actions. UI guidelines should follow the Human-Control Mantra: Preview first, select and initiate, then view execution. Similarly, UI guidelines must involve innovations that lower the barriers for users to directly influence the AI models that are supporting their tools. This could include the training regime, accessibility and inclusion and personalization. Examples include interfaces that can enable blind users to train an object recognizer with their photos or machine teaching that allows for observations and reflections and promotes user experimentation that can spark counterfactual thinking for adults and children.

3.5.2 Human-Centered Objective Functions and Datasets and Evaluation Metrics: Traditionally, objective functions had a focus on accuracy and were often brittle. More recently, they have been modified to incorporate considerations of fairness, diversity, or equity. Extensions consider preferences or value judgment aggregation techniques from the computational social choice literature. Additional consideration must be paid to avoid harm or other shortcomings, e.g., when language models only recognize binary values for human gender. They can also be extended along a dimension of noise or a surprise element, to prioritize aspects of creativity. Objective functions may also need to be tailored to specific environments, e.g., large-scale collaborative innovation platforms.

Datasets must be constructed with human-centered values in their design. A first step is to include a taxonomy to describe features or limits of the dataset. There has been interest in the past in sourcing datasets from under-represented communities, persons with disabilities, etc. It is important to make sure that technical, legal, and institutional privacy frameworks are also developed in parallel. For instance, data sourced from people with disabilities may include distinct data patterns that may be more susceptible to data abuse and misuse, e.g. risks of inaccurate or non-consenting disclosure of a disability.

References

[1] National Academies. Fostering responsible computing research: Foundations and practices. National Academies Press, 2022.

[2] Thomas Powers and Jean-Gabriel Ganascia. The ethics of the ethics of AI. In Markus Dubber, Frank Pasquale, and Sunit Das, editors, The Oxford Handbook of Ethics of AI, pages 34–39. Oxford University Press, 2020.

[3] AI risk management framework. 2022. https://www.nist.gov/itl/ai-risk-managementframework.

[4] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1:389–399, 2019.

[5] Thilo Hagendorff. The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30:99–120, 2020.

[6] Brent Mittelstadt. Ai ethics – too principled to fail? SSRN Electronic Journal, 2019.

[7] Nicholas Diakopoulos. Transparency. In Markus D. Dubber, Frank Pasquale, and Sunit Das, editors, The Oxford Handbook of Ethics of AI, pages 197–214. Oxford University Press, 2020.

[8] Alan Winfield, Serena Booth, Louise Dennis, Takashi Egawa, Helen Hastie, Naomi Jacobs, Roderick Muttram, Joanna Olszewska, Fahimeh Rajabiyazdi, Andreas Theodorou, Mark Underwood, Robert Wortham, and Eleanor Watson. IEEE P7001: A proposed standard on transparency. Frontiers in Robotics and AI, 8, 2021.

[9] Batya Friedman and Helen Nissenbaum. Bias in computer systems. ACM Transactions on Information Systems, 14(3):330–347, 1996.

[10] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 63(5):82–89, 2020.

[11] Kyla Chasalow and Karen Levy. Representativeness in statistics, politics, and machine learning. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 77–89, 2021.

[12] Selen Bozkurt, Eli Cahan, Martin Seneviratne, Ran Sun, Juan Lossio-Ventura, John Ioannidis, and Tina Hernandez-Boussard. Reporting of demographic data and representativeness in machine learning models using electronic health records. Journal of the American Medical Informatics Association, 27(12):1878–1884, 2020.

[13] David Gunning and David Aha. DARPA's explainable artificial intelligence (xai) program. AI Magazine, 40(2):44–58, 2019.

[14] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267:1–38, 2019.

[15] Maranke Wieringa. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, page 1–18, New York, NY, USA, 2020.

[16] Ben Shneiderman. Human-centered ai. Oxford University Press, 2022.

[17] Trevor Bench-Capon, Gwen Robinson, Tom Routen, and Marek Sergot. Logic programming for large scale applications in law: a formalization of supplementary benefit legislation. In Proceedings of the First International Conference on Artificial Intelligence and Law (ICAIL-87), pages 190–198. The Association for Computing Machinery Press, 1987.

[18] Marek Sergot, Fariba Sadri, Robert Kowalski, Frank Kriwaczek, Peter Hammond, and Therese Cory. The British Nationality Act as a logic program. Communications of the Association for Computing Machinery, 29:370–386, 1986.

[19] Jean-Gabriel Ganascia. Modeling ethical rules of lying with answer set programming. Ethics and Information Technology, 9:39–47, 2007.

[20] Luis Moniz Pereira and Ari Saptawijaya. Programming machine ethics. Springer, 2016.

[21] Ronald Arkin. Governing lethal behavior in autonomous robots. Chapman and Hall/CRC, 2009.

[22] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems, 77:1–14, 2016.

[23] David Abel, James MacGlashan, and Michael Littman. Reinforcement learning as a framework for ethical decision making. In Blai Bonet, Sven Koenig, Benjamin Kuipers, Illah Nourbakhsh, Stuart Russell, Moshe Vardi, and Toby Walsh, editors, AI, Ethics, and Society: Papers from the 2016 AAAI Workshop. AAAI Press, 2016.

[24] Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. AI Magazine, 36(4):105–114, 2015.

[25] Mark Riedl and Brent Harrison. Using stories to teach human values to artificial agents. In Proceedings of the 2nd International Workshop on AI, Ethics and Society, 2016.

[26] Michael Anderson and Susan Leigh Anderson. Machine ethics: creating an ethical intelligent agent. AI Magazine, 28:15–26, 2007.

[27] Michael Anderson and Susan Leigh Anderson. Geneth: a general ethical dilemma analyzer. Paladyn, Journal of Behavioral Robotics, 9:337–357, 2018.

[28] Walter Sinnot-Armstrong and Joshua August Skorburg. How AI can aid bioethics. Journal of Practical Ethics, 9, 2021.

[29] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. Toward a general logicist methodology for engineering ethically correct robots. 21(4):38–44, 2006.

[30] Formal verification of ethical properties in multiagent systems. In Proceedings of the First Workshop on Ethics in the Design of Intelligent Agents, page 26–31, The Hague, Netherlands.

[31] Dennis Walter, Holger Täubig, and Christoph Lüth. Experiences in applying formal verification in robotics. In Erwin Schoitsch, editor, Computer Safety, Reliability, and Security, pages 347–360, Berlin, Heidelberg, 2010.

[32] Alan Winfield, Christian Blum, and Wenguo Liu. Towards an ethical robot: Internal models, consequences and ethical action selection. In Michael Mistry, Aleš Leonardis, Mark Witkowski, and Chris Melhuish, editors, Advances in Autonomous Robotics Systems, pages 85–96, Cham, 2014.

[33] Matt Webster, David Western, Dejanira Araiza-Illan, Clare Dixon, Kerstin Eder, Michael Fisher, and Anthony G Pipe. A corroborative approach to verification and validation of human–robot teams. The International Journal of Robotics Research, 39(1):73–99, 2020.

[34] Supporting Human Flourishing by Ensuring Human Involvement in AI-Infused Systems, 2021.

[35] Kashmir Hill. Volvo says horrible ’self-parking car accident’ happened because driver didn’t have ’pedestrian detection’. https://splinternews.com/volvo-says-horrible-selfparking-

car-accident-happened-1793847943.

[36] Carlton Reid. Semi-autonomous cars hit cyclist in 5 out of 15 test runs,

aaa. https://www.forbes.com/sites/carltonreid/2022/05/16/semi-autonomous-car-hitcyclist-

in-5-out-of-15-test-runs-finds-aaa/.

[37] Cody Godwin. Self-driving tesla crashes into $ 3.5 million private jet using ’smart summon’

feature. https://www.usatoday.com/videos/news/have-youseen/

2022/04/25/tesla-collides-private-jet-while-owner-using-smart-summonmode/

7439216001.

[38] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. Proceedings of the National Academy of Sciences, 118(9):e2023301118, 2021.

[39] Brian Contreras. How instagram and tiktok prey on pregnant women's worst fears. https://www.latimes.com/business/technology/story/2022-05-25/for-pregnantwomen-

the-internet-can-be-a-nightmare.

[40] Anna Hill and Lamaja Denman. Adolescent self esteem and instagram: An examination of posting behavior. Concordia Journal of Communication Research, 3(1):4, 2016.

[41] Francesca Gioia, Mark D Griffiths, and Valentina Boursier. Adolescents' body shame and social networking sites: The mediating effect of body image control in photos. Sex Roles, 83(11):773–785, 2020.

[42] Torsten Kracht, Lisa Sotto, and Bennett Sooy. Facebook pivots from facial recognition system following biometric privacy suit.

https://www.reuters.com/legal/legalindustry/facebook-pivots-facial-recognitionsystem-

following-biometric-privacy-suit-2022-01-26/.

[43] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4):681–694, 2020.

[44] Amy Tennery and Gina Cherelus. Microsoft's AI twitter bot goes dark after racist, sexist tweets. https://www.reuters.com/article/us-microsoft-twitter-botidUSKCN0WQ2LA.

[45] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pages 298–306, 2021.

[46] Ryan Daws. Medical chatbot using openai's gpt-3 told a fake patient to kill themselves. https://www.artificialintelligence-news.com/2020/10/28/medicalchatbot-

openai-gpt3-patient-kill-themselves.

[47] Kadija Ferryman and Mikaela Pitcan. Fairness in precision medicine. Data & Society, 1, 2018.

[48] Ravi Parikh, Stephanie Teeple, and Amol Navathe. Addressing bias in artificial intelligence in health care. Jama, 322(24):2377–2378, 2019.

[49] Sharona Hoffman and Andy Podgurski. Artificial intelligence and discrimination in health care. Yale Journal of Health Policy & Ethics, 19:1, 2019.

[50] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modeling study. The Lancet Digital Health, 2022.

[51] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464):447–453, 2019.

[52] Jay Pujara et al. Filling the private firm void: Using learned representations to identify competitor relationships between businesses. Technical report, 2021.

[53] FEIII. The financial entity identification and information integration challenge: 2016- 2019.

2021.

[54] Wilkinson M., Dumontier M., Aalbersberg I., Appleton G., Axton M., Baak A., Blomberg N., Boiten J., da Silva Santos L., Bourne P., Bouwman J., Brookes A., Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C., Finkers R., Gonzalez-Beltran A., Gray A., Groth P., Goble C., Grethe J., Heringa J., 't Hoen P., Hooft R., Kuhn T., Kok R., Kok J., Lusher S., Martone M., Mons A., Packer A., Persson B., Rocca-Serra P., Roos M., van Schaik R., Sansone S., Schultes E., Sengstag T., Slater T., Strawn G., Swertz M., Thompson M., van der Lei J., van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J., Mons B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Nature Scientific Data, 3:160018, doi: 10.1038/sdata.2016.18, 2016.

[55] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. Language models are Few-Shot Learners. Advances in Neural Information Processing Systems, volume 33, pages 1877-1901, 2020.

[55] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Inioluwa, D., Gebru, T. Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229, 2019. https://doi.org/10.1145/3287560.3287596

[56] Gebru, T, Morgenstern, J., Vecchione, B., Vaughan, J., Wallach, Daumé III, H., and Crawford, K. Datasheets for Datasets. CoRR abs/1803.09010, 2018. http://arxiv.org/abs/1803.09010

[57] Globus-Harris, I., Kearns, M. and Roth, A. An Algorithmic Framework for Bias Bounties. 2022. https://doi.org/10.48550/arxiv.2201.10408

[58] Liu, J., Hallinan, S., Lu, X., He, P., Welleck, S., Hajishirzi, H. and CHhoi, Y. Rainier: Reinforced Knowledge Introspector for Commonsense Question Answering. 2022. arXiv:2210.03078

[62] Morgan Carlile, Brian Hurt, Albert Hsiao, Michael Hogarth, Christopher Longhurst, and Christian Dameff. Deployment of artificial intelligence for radiographic diagnosis of COVID-19 pneumonia in the emergency department. Journal of the American College of Emergency Physicians Open, 1(6):1459–1464, 2020.

[63] Brian Hurt, Meagan A Rubel, Evan M Masutani, Kathleen Jacobs, Lewis Hahn, Michael Horowitz, Seth Kligerman, and Albert Hsiao. Radiologist-supervised Transfer Learning. Journal of Thoracic Imaging, (00):1–10, Oct 2021.

[64] Justin Huynh, Samira Masoudi, Abraham Noorbakhsh, Amin Mahmoodi, Seth Kligerman, Andrew Yen, Kathleen Jacobs, Lewis Hahn, Kyle Hasenstab, and Michael Pazzani. Deep learning radiographic assessment of pulmonary edema: Training with serum biomarkers. IEEE Access, 2022.

[65] Michael Pazzani, Robert Kaufman Severine Soltani, Samson Qian, and Albert Hsiao. Expert-informed, user-centric explanations for machine learning. In Proceedings of the AAAI Conference on Artificial Intelligence-2022. IOS Press, 2022.

[66] Samson Qian. Generating explanations for chest medical scan pneumonia predictions. COVID Information Commons, 2021.

Authors
Kevin Ashley, University of Pittsburgh
Iliaria Canavotto, University of Maryland
Jill Cristman, UL Research Institute
David Danks, University of California, San Diego
Kristian Hammond, Northwestern University
John Horty, University of Maryland
Michael Pazzani, USC/ISI
Louiqa Raschid, University of Maryland
Daniel Schiff, Purdue University and JP Morgan Chase

Appendices

A1: Workshout Breakout Group Reports on Themes

A1.1 Best Practices

What are the key challenges that should be addressed within this theme? What are the research questions that would be appropriate for these challenges?

- How can we apply the lessons that have been learned from previous technological revolutions into the ethical design of AIs? How can we avoid reinventing various wheels? How might we enable people to learn what worked (ethically) and what did not work?
- How can we develop inclusive design protocols that engage all stakeholders?
- We need a systems engineering process that supports the incorporation of ethical design practices, from requirements gathering all the way through to deployment and maintenance.
- How do we develop co-design methods that can enable the precise elicitation of goals for the design of ethical AIs?
- Best practices to address data quality and privacy challenges:
    - How do we assure that AIs do not have more errors on minority classes than on the majority class?
    - How might we connect data sources when some or all of the data are related to protected populations?
    - How do we encourage increasing representativeness in data sets?
    - How do we identify and "complete" data gaps?
    - How do we determine that a given model / dataset is 'fit' for use?
    - How can we support transparency about what variables or outcomes are being given priority?
- How do we become proactive, e.g., predicting threats and developing community standards to guard against harm, prior to the deployment of AIs?
- How do we design AIs that are incentivized to limit gaming the system? How can we encourage users / actors to provide accurate and unbiased information?
- Can we create software libraries / frameworks that have safety "built-in"?
- How do we identify high risk AI implementations that might cause harm to society as a whole?
- How do we establish best practices for the development of systems or pipelines with AI components, beyond individual AIs?
- Many AI pipelines may include a need for a "human in the loop". Where in the loop do we inject the human? Can we inject teams?
- How do we identify scenarios where an approach, e.g., crowd-sourcing, might introduce bias? What tools can be developed to help uncover problems in bias?
- How do we make sure that AIs do not reproduce or replicate traditional stereotypes, e.g., a female helper such as Siri or Alexa, whereas online medical diagnostic systems often present as males? We need to recognize that gender and presentation and identification go beyond voice or skin tone.
- How can we increase the opportunities or reduce the barriers to deploy AIs for beneficial purposes? A caveat that this may simultaneously facilitate abuse and misuse.

A1.2 Ethical AI Governance

What are the key challenges that should be addressed within this theme? What are the research questions that would be appropriate for these challenges?

- How can one incentivize companies to develop (and use) best practices, starting early in the design phase? This is in contrast to the more common first-to-market incentives.
    - Can we provide tools that are easily available and easy to use, and that can help to produce AIs that are, for example, more equitable?
    - Can we provide financial incentives to create open-source tools and open data for training and testing?
    - Can we ensure that the costs / benefits are (equally) shared across all stakeholders, and that no one group is at a disadvantage?
- How do we incentivize preventive / proactive approaches and self-governance over reactionary punishment?
- How do you create a policy framework to support / promote open data, explicit metadata to capture data provenance, transparency around specific design decisions, etc.?
- What approaches can be used to take the burden of proof off the people who experience harm, both in specific cases and systemically?
- How can we develop methods to translate norm-based governance into relevant features? How can we check that the AI accurately captures the norms that are required by the law or are important to stakeholders?
- It is very hard to operationalize values in a domain-agnostic manner. How do you craft regulations and other governance mechanisms - that seek to establish "standards" - to be both domain relevant and more widely applicable?
- How can you construct a governance mechanism to track changes? This could include technology evolution and changes to stakeholder requirements.
- What public/community/open resources can we make available to organizations, in particular, NGOs or nonprofits, for a range of tasks? The tasks can be very focused, e.g., identifying bias in the outcomes of some AIs, to broad brush, e.g., determining if a company or product has performed due diligence / duty of care with respect to the design or deployment of some AI.
- How does one develop standards for accountability?
    - Who is liable for harms that occur during the development of AIs?
    - Where does the responsibility of the AI developer end?

A1.3 Measures

What are the key challenges that should be addressed within this theme? What are the research questions that would be appropriate for these challenges?

- What is a formal (quantitative or qualitative) representation of a goal (measure) for ethical AI? [We use the terms goal and measure somewhat interchangeably.]
- Systems often have multiple goals. Can we quantify the trade-off between these goals? Can we capture these trade-offs using some measure?
- Tradeoffs may depend critically on the requirements and values of specific communities or problem domains. This makes it unlikely to produce a one-size-fits-all measure to study tradeoffs.
- How do we identify broad (background) assumptions that an AI is expected to meet? How do we measure or evaluate the impact(s) of not meeting those expectations or of potential violations of those assumptions?
- What are some historical examples of safety evaluations (or similar) across a range of application domains (not limited to AIs)? What principles and frameworks and lessons can be translated to AIs?
- What foundational changes in measurement theory are required to allow for dynamic changes in goals, measures and AIs? Can we build on existing work that allows temporal changes in measures?
- How can we support the validation and verification of goals and measures?
    - What are the appropriate approaches for validation within an organization and / or for independent third parties?
    - There is a need for iterative refinement and validation, starting at the early stages of the design of the AI.

A1.4 Values

What are the key challenges that should be addressed within this theme? What are the research questions that would be appropriate for these challenges?

- How do we apply democratic and participatory policy during the design of AI systems?
- How do we overcome roadblocks like the lack of public understanding of AI systems?
- What frameworks for values elicitation are appropriate, and for what contexts?
- How do we formalize the process of value elicitation from stakeholders, contexts, cultures, and normative frameworks?
- Who decides which values are important? How should values that evolve over time be addressed?
- How do we operationalize values in AI systems?
- How do we (designers, stakeholders, governance bodies) navigate trade-offs between values? How do we navigate trade-offs between humans and AIs?
- How do we weigh values - privacy vs. benefit; fairness vs. benefit - when there are less clear tradeoffs?
- How can we (or the AI) determine the appropriate context to frame an ethical decision?

If we could solve these challenges, what would be the positive societal outcomes?

- A common framework with which we can evaluate AI ethics debates, design processes, and build systems.
- Multi-stakeholder agreement, in the broadest sense, and in particular, an increased engagement of impacted communities in the ethical AI design process.
- Improved trust between communities and AI systems (and developers).
- Improved awareness of, and education about, AI systems.
- ‘Who designs AI’ becomes more open and more transparent.
- AI doesn’t become the next automobile, i.e., yielding some benefits but eventually leaving a disastrous impact.
- Increased clarity in the specification of ethical values can lead to more AI research achievements a la the positive impact of counterfactuals advancing machine learning.
- A reevaluation of current/future systems to improve their match to stakeholder values.
- Enhanced creativity and innovation due to the embrace of a multi-disciplinary, multi-stakeholder approach.
- A successful launch of an “Ethical Values by Design” standard for use across industry and academia.

A1.5 Organizations and Incentives

What are the key challenges that should be addressed within this theme? What are the research questions that would be appropriate for these challenges?

- How do we know if / when ethical activities are making a difference? How do we measure progress and understand if incentives and organizational structures are working?
- How do we incentivize practitioners to care about notions beyond test data performance, e.g., safety? Metrics such as accuracy, precision, and area under the curve are easy to quantify and to show improvement. How do we incentivize safety especially since it is not easy to quantify or optimize safety around a safety metric.
- How do we create incentives for “voluntary” external auditing and evaluation, perhaps by independent third parties? Should we recommend the use of more transparent models, in comparison to more opaque models, to identify cases for audit?
- How do we make sure that interdisciplinary teams, e.g., HCI, ethics, and social scientists are involved at all stages of the development and not included as an afterthought?
- How do we encourage sharing and integrating data sets? How to address the problem of a company saying "I can't show you my data because it's proprietary"? We need mechanisms for data sharing that can allow for testing and validation without the need to disclose entire datasets.
- When it comes to the design of incentives, we should incorporate design research on economic + experimental and behavioral + experimental design.
- What specific organizational structures affect change? Boards? Risk committees? Lead AI / ML Director? Where do they sit in org charts for most effectiveness?
- What types of incentives actually work? This may include economic and non-economic incentives, e.g., behavioral, brand / reputation, public service motivation, etc.
- What types of interdisciplinary mechanisms / teams work? How can they be incentivized to drive real world outcomes? What are the appropriate roles of different disciplines? What kinds of practices should be shared or separated, and in what organizational structures and processes? How do they fit into org structures and project management best practices?
- How do we make any of this work for smaller organizations without deep pockets and many people?

If we could solve these challenges, what would be the positive societal outcomes?

- Increased safety and AI that mitigates / eliminates vs. reinforces existing inequities and disparities.
- Reduced quantity / impact of AI incidents (failures, abuses, misuses, etc.).
- Broader adoption or benefit of AI/ML technologies, particularly those not directly for commercial gain, e.g., in government, civil society, non-profit, education, etc.
- Improved regulations that permit technological progress while avoiding negative consequences.
- AN understanding of what incentives and organizational practices should be taught in educational settings.
- Guidelines for the robust interdisciplinary practice of safe and ethical AI.
- Professional development for “Responsible AI” career paths.

A1.6 Training and Education

What are the key challenges that should be addressed within this theme?

- Educating the public and / or users about the benefits and the potentially harmful limitations of AIs.
- Leveraging lessons learned from other domains or historical inventions, to build an understanding of current AI technology & best communication practices.
- Training technologists in ethics and training ethicists about technology.
- Providing the relevant training for non-technologists who are professionals in the law, regulation and compliance, or in domains in which AIs are extensively deployed, so they can contribute meaningfully to ensuring positive outcomes and minimizing harm.
- Educating human-centered designers and users on the need to address the landscape after the successful deployment of an AI.

What are the research questions that would be appropriate for these challenges?

- What specific information do technologists across the AI lifecycle need to know about ethics in order to create ethical AI products/outcomes? What are the best methods to disseminate this information? What are some specific tools that will aid them?
- What previous lessons in building public understanding or public acceptance from other past technologies or domains can be incorporated when we think about societal adoption of AI? [Note: Public understanding does not equal public acceptance.]
- How can we balance providing members of the public the right information about AI's limitations with also encouraging continued use in domains where AI has great potential?
- How can one design AI systems that anticipate and monitor improper use of the system?
- How can one build incentives into an AI system to encourage its proper use and avoid gaming the system?
- How can education and training of users and designers help to ensure that the AI system is used properly?

What are the obstacles/hurdles/needs within these challenges that must be satisfied in order for you to succeed?

- Clear terminology that can be expressed across disciplines and remain accessible to public audiences outside of academia
- Easily understandable metrics for assessing ethical standing for an AI system
- Lack of educational best practices/modules/specific actionable information that can be taken to the public
- Need for different discipline communities to understand each other's methodologies

Recommendations for Training and Education:

- Develop case studies that illustrate best practices.
- Develop educational courses / materials for technologists to learn standardized information about ethics.
    - Could piece together elements from different courses that already exist in some institutions.
    - Could be targeted for students receiving CS/AI/END degrees or targeted at career level professionals.
- Develop a prototype of an educational course/materials for ethicists and legal experts to

learn standardized information about AI technology.

- Develop a prototype of an educational course/materials for ethicists and legal experts to learn standardized information about AI technology.
    - Pros: could help cut through the hype and misinformation about AIs.
    - Cons: might be difficult to keep up to date as the technology changes/evolves.
-
- Develop an educational workshop with both tech developers and ethicists to discuss developing an AI model and navigating potential ethical issues during the design process.
- Develop an ongoing training across the AI lifecycle for users who weren't involved with the original development.

A2: NSF EDAIs Workshop Schedule

September 22 2022

- 1:00 PM EDT Overview of the program
- 1:20 PM EDT Maja Mataric, USC (Introduced by Ilaria Canavotto)
- 1:40 PM EDT Michael Kearns, Penn (Introduced by John Horty)
- 2:00 PM EDT Discussion (Led by Jim Hendler)
- 2:10 PM EDT Ece Kamar, Microsoft (Introduced by Ryan Jenkins)
- 2:30 PM EDT Molly Steenson, CMU (Introduced by Daniel Schiff)
- 2:50 PM EDT Discussion (Led by Leora Morgenstern)
- 3:00 PM EDT Survey Discussion (Michael Pazzani)
- 3:10 PM EDT Wrap up discussion on EDAIs Challenges (David Danks)

September 29 2022

- 12:00 pm Welcome
- 12:25 pm Breakout Activity
- 01:05 pm Breakout Activity
- 01:45 pm Break
- 02:10 pm Theming
- 02:40 pm What's Missing?
- 03:00 pm Provocateur: Oren Etzioni, Allen Institute for Artificial Intelligence
- 03:15 pm Breakout Activity

October 06 2022

- 12:00 pm Welcome
- 12:10 pm Clustering Challenges
- 12:35 pm Reflect on the Themes / Break
- 01:05 pm What's missing & voting
- 01:25 pm Break
- 01:40 pm Sign-up & Breakout time
- 02:45 pm Report Back with Feedback
- 03:25 pm Closing & Next Steps

October 20 2022

12:00 PM Presentation - "Why Do Ethical AI?" David Danks

12:20 PM Industry Panel - "Real User Needs"

- Diane Staheli, Chief, Responsible AI, US Department of Defense Chief Digital and AI Office
- Aruna Rajan, Director of Applied ML, Google India
- Erica Smith, Unit Chief, Bureau of Justice Statistics
- Mona Diab, Lead for Responsible AI, Meta

1:00 PM Breakouts Round 1

1:15 PM Report Back

1:50 PM Breakout Round 2

2:20 PM Closing Statements

Appendix A3: Subset of NSF grants related to AI & Ethics

PI	Organization	Title
David Benkeser	Emory University	Accurate and Interpretable Machine Learning for Prediction and Precision Medicine
H Jagadish	University of Michigan - Ann Arbor	BIGDATA: F: Collaborative Research: Foundations of Responsible Data Management
Yulia Tsvetkov	University of Washington	CAREER: Language Technologies Against the Language of Social Discrimination
Renran Tian	Indiana University	CAREER: Modeling Situated Intention during Nondeterministic Pedestrian-Vehicle Interactions through Explainable Compositional Learning of Naturalistic Driving Data
Alan Wagner	Pennsylvania State Univ	CAREER: No Time to Explain: Developing Robots that Actively Prevent Overtrust during Emergencies
Olga Russakovsky	Princeton University	CAREER: Overcoming bias in computer vision: Building fairer systems and training diverse leaders
Giuseppe Loianno	New York University	CAREER: Re-Thinking the Perception-Action Paradigm for Agile Autonomous Robots
Hadi Hosseini	Pennsylvania State Univ	CAREER: Robust Fairness in Matching Markets
Peng Wei	George Washington University	CAREER: Safe and Scalable Learning-based Control for Autonomous Air Mobility
Casey Fiesler	University of Colorado at Boulder	CAREER: Scaffolding Ethical Speculation in Technology Design
Christopher Dancy	Pennsylvania State Univ	CAREER: SocioCulturally Competent Agents to Study and Improve Human-AI interaction
Diyi Yang	Georgia Tech	CCRI: Research Infrastructure: Planning-M: Multi-Modal Infrastructure for Enabling Social AI Research
Arvind Narayanan	Princeton University	CHS: Large: Collaborative Research: Pervasive Data Ethics for Computational Research

Mor Naaman	Cornell University	CHS: Medium: Collaborative Research: Charting a Research Agenda in Artificial Intelligence-Mediated Communication
Colin Gray	Purdue University	CHS: Small: Improving Everyday Ethics in Socio-technical Practice
Shandong Wu	University of Pittsburgh	CICI: SIVD: Discover and defend cyber vulnerabilities of deep learning medical diagnosis models to adversarial attacks
Ruth West	University of North Texas	Collaborative Research: NRI: FND: Grounded Reasoning about Robot Capabilities for Law and Policy
Deirdre Mulligan	University of California Berkeley	Collaborative Research: Standard: Emerging Cultures of Data Science Ethics in the Academy and Industry
Cathryn Carson	University of California Berkeley	Convergence HDR: Social Science Insights for 21st Century Data Science Education (SSI)
Chao Lan	University of Oklahoma	CRII: III: Fair Machine Learning with Restricted Access to Sensitive Personal Data
Prabha Sundaravadivel	University of Texas at Tyler	CyberTraining: Implementation: Small: Collaborative Research: Easy-Med: Interdisciplinary Training in Security, Privacy-Assured Internet of Medical Things
Saraju Mohanty	University of North Texas	CyberTraining: Implementation: Small: Collaborative Research: Easy-Med: Interdisciplinary Training in Security, Privacy-Assured Internet of Medical Things
Sanmay Das	George Mason	EAGER: AI-DCL: Exploratory research on the use of AI at the intersection of homelessness and child maltreatment
Robin Murphy	Texas A&M	EAGER: Evidence-Based Model of Adoption of Robotics for Pandemics and Natural Disasters
Michael Anderson	University of Hartford	EAGER: Toward Ethical Intelligent Autonomous Systems, A Case-Supported Principle-Based Behavior Paradigm
Jennifer Jacobs	University of California-Santa Barbara	Ethical and Responsible Research for Augmented Reality
Nathan Kallus	Cornell	FAI: Auditing and Ensuring Fairness in Hard-to-Identify

	University	Settings
Shiri Dori-Hacohen	University of Connecticut	FAI: BRIMI - Bias Reduction In Medical Information
Jiang Li	Howard University	HDR DSC: Collaborative Research: Transforming Data Science Education through a Portable and Sustainable Anthropocentric Data Analytics for Community Enrichment Program
Yu Liang	University of Tennessee Chattanooga	HDR DSC: Collaborative Research: Transforming Data Science Education through a Portable and Sustainable Anthropocentric Data Analytics for Community Enrichment Program
Pablo Rivas	Baylor University	IUCRC Planning Grant Baylor University: Center for Standards and Ethics in Artificial Intelligence (CSEAI)
Junfeng Jiao	University of Texas at Austin	NRT-AI: Convergent, Responsible, and Ethical Artificial Intelligence Training Experience for Roboticists
Amy Pruden	Virginia Polytechnic Institute &SU	NRT-HDR: Convergence at the Interfaces of Policy, Data Science, Environmental Science and Engineering to Combat the Spread of Antibiotic Resistance
Trisha Phillips	West Virginia University	RAPID: Using a professional code of ethics to promote ethical and responsible research
Arvind Narayanan	Princeton University	RI: Medium: Recognizing, Mitigating and Governing Bias in AI
Munindar Singh	North Carolina State University	RI: Small: Foundations of Ethics for Multiagent Systems
Veronica Ahumada-Newhart	University of California-Davis	Robot-Mediated Learning: Exploring School-Deployed Collaborative Robots for Homebound Children
Thomas Williams	Colorado School of Mines	S&AS: FND: Context-Aware Ethical Autonomy for Language Capable Robots
Shlomo Zilberstein	University of Massachusetts Amherst	S&AS: FND: Reliable Semi-Autonomy with Diminishing Reliance on Humans
Alan Wagner	Pennsylvania State Univ	S&AS: INT: COLLAB: Do the Right Thing: Competing Ethical Frameworks Mediated by Moral Emotions in

		Human Robot Interaction
Ronald Arkin	Georgia Tech	S&AS:INT:COLLAB:Do the Right Thing: Competing Ethical Frameworks Mediated by Moral Emotions in Human-robot Interaction
Bimal Nepal	Texas A&M	Standard Research: Developing Ethical STEM Research Competency and Self-Efficacy in High School and College Engineering Courses
Karen Levy	Cornell University	Standard: Collaborative Research: Emerging Cultures of Data Science Ethics in the Academy and Industry
Tom Yeh	University of Colorado at Boulder	STEM+C: Integrating AI Ethics into Robotics Learning Experiences
Kristen Venable	University of West Florida	TRAVEL PROPOSAL: STUDENT PROGRAM OF THE FIFTH CONFERENCE ON AI, ETHICS AND SOCIETY (AIES 2022)

Appendix A4: Workshop Participants

Michael Anderson, University of Hartford
Kevin Ashley, University of Pittsburgh
Solon Barocas, Microsoft Research
Jean Camp, Indiana University
Tabitha Colter , MITRE
Sanmay Das , George Mason University
Huiling Ding , NC State University
Shiri Dori-Hacohen , University of Connecticut + AuCoDe
Kadija Ferryman, Johns Hopkins University
Juliana Freire, New York University
Ashok Goel , Georgia Tech
Colin Gray , Purdue University
Cindy Grimm, Oregon State University
Swati Gupta , Georgia Institute of Technology
Patrick Hall
Kristian Hammond, Northwestern University
John Hearty , Mastercard
Hadi Hosseini, Penn State University
Dave Kaufman , Georgetown University
Ramayya Krishnan , CMU
Chao Lan , University of Oklahoma
Cara LaPointe, Johns Hopkins Institute for Assured Autonomy
Derek Leben , Tepper School of Business, Carnegie Mellon University
Nicholas Mattei, Tulane University
Mark Nitzberg, Center for Human-Compatible AI (CHAI)
Lynne Parker, University of Tennessee, Knoxville
Aruna Rajan, Google
Aaron Roth , Naval Research Laboratory, and University of Maryland
Stuart Russell, UC Berkeley
Raesetje Sefala, DAIR Institute
Olivia Sheng, University of Utah
Katie Shilton, University of Maryland, College Park
David Shmoys, Cornell University
Munindar Singh, NCSU
Erica Smith, Bureau of Justice Statistics
Brittany Smith, Schmidt Futures
Diane Staheli, US Department of Defense, Chief Digital and AI Office

Renran Tian, Indiana University Purdue University Indianapolis
Phebe Vayanos, University of Southern California
Ruth West, University of North Texas
Shandong Wu, University of Pittsburgh
Lirong Xia, RPI
Holly Yanco, University of Massachusetts Lowell
Shlomo Zilberstein, Stanford University