Response to NIST RFI on AI Executive Order 14110:

The Importance of a Socio-technical Approach in AI Development

Submitted by the Institute for Trustworthy AI in Law and Society

February 2, 2024

I. Introduction

The National Institute of Standards and Technology (NIST) plays a crucial role in guiding the development and deployment of Artificial Intelligence (AI) systems as mandated by Executive Order 14110. As AI technologies become increasingly integrated into various sectors of society, it is imperative to adopt a ***socio-technical*** approach in creating guidelines, standards, and best practices for AI safety, security, and trustworthiness.

The National Institute of Standards and Technology's (NIST) Request for Information (RFI) under the Executive Order on AI represents a pivotal opportunity for TRAILS (The NIST-NSF Institute for Trustworthy AI in Law & Society) to offer its unique insights. Through integrating AI technology, participation, and governance, TRAILS is reshaping AI practices towards practical, ethical, human rights-centered paradigms. Supported by the National Science Foundation and NIST, in collaboration with major universities and industry players, TRAILS focuses on developing AI systems that are trustworthy, accountable, and reflective of diverse stakeholder perspectives. Because our mission is to ensure that AI systems not only enhance human capabilities but also uphold human dignity and rights, we emphasize methods that bolster AI trustworthiness, user empowerment, and inclusive governance. Through multidisciplinary research and training, TRAILS is committed to bringing forward voices often overlooked in mainstream AI development, ensuring a comprehensive and inclusive approach to AI governance.

This approach is essential for formulating guidelines, standards, and best practices for AI safety, security, and trustworthiness that achieve the necessary, intricate balance between technological advancement and social impact. Therefore, TRAILS' mission and expertise powerfully aligns with NIST's crucial role in guiding AI development and deployment. To ensure the safety AI technologies, as they become more prevalent across various societal sectors, NIST should adopt a socio-technical framework. A socio-technical approach allows AI governance to evolve as AI related systems change over time.

Extensive research conducted by TRAILS researchers demonstrates the necessity of a socio-technical approach to AI. This approach considers both the social and behavioral context alongside the technological aspects of AI. This broader persepctive is critical for creating systems that are not only technologically advanced but also socially responsible and ethically sound. By integrating insights from diverse fields such as law, social sciences, and computer science, TRAILS research underscores the importance of addressing the multifaceted impacts of AI on society. This comprehensive perspective ensures that AI systems are developed with a keen awareness of their potential societal implications, aligning with our shared goals of trustworthiness, accountability, and inclusivity.

II. What is a Socio-Technical Approach?

A socio-technical approach comprehensively examines the interaction between technology, how it is used in practice, and the consequences of this use. This methodology extends beyond examining the technical aspects of systems, also focusing on how technology integrates and interacts with psychological, societal, economic, and ethical dimensions.

As technology becomes more powerful and adoption grows, it increasingly influences, and is influenced by, human behavior (including individual, social, cultural and regulatory considerations). An approach that takes these interactions into account is crucial in the context of AI. The impact of AI extends far beyond its immediate functionality, affecting job markets, legal frameworks, privacy norms, and ethical boundaries. By considering these broader impacts, taking a socio-technical approach reduces the risk that AI development works against the values and needs of the people it affects and society as a whole. It aims to create systems that are not only efficient and effective but also respect social and ethical imperatives, such as by examining how the benefits of AI are distributed and how its risks can be mitigated.

One might think that such an approach sacrifices technological innovation to promote social well-being. However, ***technological innovation cannot be effective if technology is not adopted***. A lack of trust is a major reason technology might not be adopted. Trust forms a crucial aspect of social capital and is integral to the successful integration of AI in society. Building and establishing trust in AI is fundamentally a socio-technical challenge, as it involves not only the technical performance of AI systems but also how they are perceived. Do users of these technologies, as well as those who make decisions about their adoption and application, experience them as beneficial, ethical, and aligned with their values? The socio-technical approach thus emphasizes the dual need to advance technology while concurrently nurturing trust among users and stakeholders.

III. AI at a Tipping Point: The Need for a Socio-technical Approach

The incorporation of a socio-technical approach in AI is neither unique nor unusual, but rather a normal phase in the development cycle of any maturing technology. Historically, as technologies evolve and become more integrated into society, they naturally transition from purely technical domains to socio-technical realms. Adopting a socio-technical perspective is not only normal but also essential for the continued innovation and integration of technology in society. Technologies that are not aligned with societal needs and values will not be readily adopted.

The consideration of socio-technical requirements in AI is appropriate to its stage of maturation. AI has successfully crossed the threshold from a technical curiosity to a functioning technology that has become incorporated into our daily lives. The growing maturity of AI technologies is evidenced by the degree to which they already significantly influence aspects of society including healthcare, legal systems, democracy, and employment. Such broad impact can only be documented and assessed accurately through a socio-technical framework focused on safe, secure, trustworthy, and socially responsible development and deployment of AI systems. Consequently, embracing a socio-technical approach for AI is a natural, expected, and crucial step in its evolution, ensuring its continued success.

IV. Socio-technical Precedents in Other Engineering Fields

Designers and regulators of all widely adopted technologies have had to adopt socio-technical considerations when they became sufficiently mature. Historical precedents in various engineering disciplines demonstrate this consistent pattern of evolution from purely technical focus to socio-technical integration.

- In *civil engineering*, the transition from a focus on purely technical aspects like material selection (e.g., asphalt vs. concrete) to a deeper engagement with social concerns has been significant. Two examples illustrate this evolution:
    - Urban Planning and Public Health: Modern civil engineering integrates public health considerations into urban planning. For example, engineers design pedestrian-friendly cities not only with the technical aspects of road and building design in mind, but also the health benefits of encouraging walking and reducing vehicle emission. Their work is informed by an understanding how people experience and make choices about how they interact with the built environment in order to realize such benefits. This approach reflects a socio-technical perspective, balancing technical infrastructure development with public health objectives.
    - Noise Control in Urban Environments: Another example is the development of noise control regulations in urban areas. Civil engineers now consider the social impact of noise pollution when designing highways and urban spaces. This includes implementing noise barriers and designing building layouts that minimize noise impact on residential areas. These measures demonstrate a socio-technical approach where technical expertise is applied in tandem with understanding and mitigating social nuisances like noise pollution.
- In *automotive engineering*, the integration of socio-technical approaches has also been evident through the emphasis on safety features and environmental considerations. Examples include:
    - Safety Features: The incorporation of advanced safety features like airbags, anti-lock braking systems (ABS), and electronic stability control (ESC) in vehicles is a prime example. These features represent a blend of technical innovation and a deep understanding of human safety needs. The design and implementation of these systems require a balance between mechanical engineering, electronics, and the psychology of driver behavior, reflecting a socio-technical approach to vehicle safety.
    - Environmental Considerations: Automotive engineering has increasingly focused on reducing the environmental impact of vehicles. This is exemplified by the development of electric and hybrid vehicles, which are designed to minimize emissions and reduce dependency on fossil fuels. This shift not only addresses the technical challenges of developing alternative energy vehicles but also reflects a broader societal concern for environmental sustainability.

- In *aeronautical engineering*, the socio-technical approach is exemplified by the transition to stringent regulations and the incorporation of *human factors engineering*:
    - Human Factors Engineering: The field has increasingly focused on human factors engineering, which involves designing aircraft systems and cockpits in a way that accounts for human capabilities and limitations. This includes ergonomic design, intuitive controls, and systems that reduce the likelihood of human error. By considering the interaction between pilots, crew, and aircraft systems, aerospace engineering demonstrates a socio-technical approach, ensuring that technology is designed with human factors in mind to enhance overall safety and efficiency.
    - Stringent Regulations: Aerospace engineering has evolved to include rigorous regulatory standards to ensure the safety and reliability of aircraft and spacecraft that account for these human factors. This includes detailed certification processes, regular safety inspections, and adherence to international aviation standards. These regulations represent a socio-technical approach that balances technical excellence in design and manufacturing with comprehensive safety protocols and oversight, addressing not just the functionality of the aircraft but also the safety of passengers and crew.
    -
- In *crewed space systems systems engineering*, the socio-technical approach is widespread. For example:
    - Space system development: Following the largely technical imperative of the "space race" against the USSR, US space exploration efforts pivoted to emphasizing concerns pertaining to geopolitical considerations and national prestige. For example, technological choices pertaining to the Space Shuttle and International Space Station typically incorporated political considerations, including which constituency would obtain the most economic benefit from the location of different NASA centers.
    - Astronaut selection: Additionally, teams of astronauts preparing for long-term space missions routinely need to consider dynamics that are both internal to the team, such as how to minimize the potential for conflict while in a long-term high-pressure environment, and external to the team, such as how to ensure that the membership of the team is embodies the diversity of the American population, while also facilitating international collaboration.
- Although typically evaluated using "hard" metrics, *development of military weapon systems* are also strongly driven by socio-technical considerations:
    - Precision-Guided Munitions (PGMs): The development of PGMs involves not only technical expertise in guidance systems but also considerations of international law and ethical warfare. To reduce collateral damage, these weapons are designed to be more accurate than necessary simply to destroy their targets.
    - Unmanned Aerial Vehicles (UAVs) or Drones: The use of drones in military operations incorporates advanced technology for surveillance and targeting while also considering the legal and ethical implications of remote warfare, including civilian safety and international regulations.

- *Biomedical engineering* is another field that has a long history of incorporating socio-technical considerations:
    - Medical Device Development: Biomedical engineering has seen a shift towards more stringent regulations, especially in medical device development. For example, the introduction of devices like pacemakers and artificial joints requires not only technical innovation but also rigorous clinical testing and approval, techniques to ensure representativeness in clinical trials, and a robust postmarket surveillance regime.
    - Ethical Considerations in Genetic Engineering: The field of genetic engineering within biomedical engineering is a prime example of ethical considerations at play. Techniques like CRISPR for gene editing have revolutionized medicine but also raised significant ethical debates regarding potential long-term effects and the ethical implications of gene manipulation, leading some nations to put a moratorium on this technology.
- In geological engineering, there has been a significant shift towards recognizing environmental impacts and ensuring regulatory compliance. This evolution reflects a socio-technical approach that balances technical expertise with environmental stewardship and adherence to legal standards.
    - Environmental Impact Assessments in Mining: Geological engineers are now required to conduct comprehensive environmental impact assessments before starting mining projects. This includes evaluating potential effects on ecosystems, water quality, and local communities, ensuring that mining activities comply with environmental regulations and minimize ecological damage.
    - Fracking Regulations: The development and implementation of hydraulic fracturing or fracking techniques in oil and gas extraction have led to increased regulatory scrutiny. Geological engineers must adhere to strict guidelines to prevent groundwater contamination and seismic disturbances, balancing the technical aspects of resource extraction with environmental protection and public health concerns.
- In the field of communication technologies, the interplay between regulation, societal norms, and technology is evident. This socio-technical approach underscores how technological advancements are guided by social and legal frameworks.
    - FCC Regulation: The Federal Communications Commission (FCC) in the United States plays a crucial role in regulating communication technologies. This includes setting standards for broadcast content, managing radiofrequency use, and enforcing rules about indecency and profanity in broadcasting, ensuring that communication technologies align with legal standards and public interest.
    - Movie Ratings: The movie rating system, established to categorize films based on their suitability for various audiences, is another example of societal norms influencing technology. This system guides content creators and distributors in making films that are appropriate for their intended audience, balancing creative expression with societal expectations and moral standards.
- In nuclear engineering, the socio-technical approach is evident in the establishment of international safety and non-proliferation standards. This field not only encompasses

advanced scientific and technical expertise but also a deep commitment to global safety and ethical responsibilities. Examples include

- International Atomic Energy Agency (IAEA) Safeguards: The IAEA implements safeguards to monitor nuclear programs worldwide, ensuring they are used for peaceful purposes and not for the development of nuclear weapons. This involves a complex mix of technical inspections and political diplomacy.
- Nuclear Reactor Safety Standards**: The design and operation of nuclear reactors are subject to stringent international safety standards. These standards are designed to prevent accidents, minimize radiation exposure to workers and the public, and ensure safe disposal of nuclear waste, balancing technical challenges with public health and environmental protection.

The above examples demonstrate how a socio-technical approach is both widespread and can result in many different types of outcomes, depending on the complex interactions between society and technology.

V. AI as a Socio-technical System

To effectively manage risk, treating AI as a socio-technical system during safety evaluations is essential. The evolution of AI from a niche technological field to a system with significant societal impacts mirrors the development trajectory observed in fields like civil, automotive, and nuclear engineering. These fields have transitioned from focusing solely on technical aspects to integrating socio-technical considerations, acknowledging the interplay between technology and societal factors.

AI systems, like those in other mature technological fields, have far-reaching implications beyond their immediate functionality. They impact environmental sustainability, economic structures, legal frameworks, and societal norms. For instance, just as stringent regulations and human factors are vital in aerospace engineering, AI evaluations must consider ethical, legal, and social implications alongside technical metrics. For new technology to flourish, policymakers must create an enabling environment that nourishes innovation while protecting market participants and the public alike. Effective policy creates an enabling environment that builds trust and reduces uncertainty for market actors. As a report for the US National Academy of Sciences notes, the only way to govern such complex systems is to create “a governance ecosystem that cuts across sectors and disciplinary silos and solicits and addresses the concerns of many stakeholders” (Gary A Marchant and Wendell Wallach 2015; Mathews et al. 2022).

The adoption of a socio-technical approach in evaluating AI aligns with the NIST’s mandate for AI systems to be safe, secure, and trustworthy. This approach enables a holistic evaluation of AI, ensuring that systems are not only accurate and efficient but also align with societal values, ethical standards, and legal requirements. The transition to viewing AI through a socio-technical lens is a natural progression, reflecting its maturity and ubiquity in modern society. This perspective is crucial for AI to be responsibly integrated into various aspects of life, ensuring its benefits are maximized while minimizing potential harms.

For example, adopting an engineering systems approach, as fostered by organizations like the Council of Engineering Systems Universities (CESUN), is crucial for addressing the complexities of AI as a socio-technical system. CESUN, established by universities dedicated to advancing engineering systems as a field of study, emphasizes the importance of tackling large-scale, interconnected socio-technical systems. (TRAILS researcher Zoe Szajnfarber is the current of chair of CESUN.) Specifically, this tradition focuses on promoting trust in AI systems within the context of the "ilities" - characteristics such as reliability, scalability, and sustainability. These "ilities," as defined by the engineering systems movement, align closely with the prerequisites for trustworthy AI outlined in NIST's AI Risk Management Framework. The framework emphasizes the importance of ensuring that AI systems are accurate, robust, resilient, explainable, and accountable. Evaluating AI systems in terms of these "ilities" ensures that they not only function efficiently but also gain the trust of users by being valid and reliable, safe, fair (bias is managed), secure and resilient, transparent and accountable, explainable and interpretable, and privacy-enhanced. This approach, which is fundamentally socio-technical, is essential for AI evaluations in the context of the NIST RFI, as it ensures that AI systems are not only technically sound but also socially responsible and aligned with broader societal goals. An engineering systems perspective integrates multiple disciplines, focusing on complex system behaviors, interactions, and societal impacts, making it a fitting framework for evaluating the maturity and innovation of AI technologies.

VII. The importance of a participatory approach

Incorporating a participatory approach in AI development is vital to align the technology with societal norms and ethical expectations. Engaging a broad cross-section of society throughout the AI development lifecycle ensures diverse perspectives are considered. This inclusivity is crucial for anticipating and addressing social implications, preventing the exacerbation of existing inequalities, and resolving tensions. A participatory approach fosters transparency and public trust in AI systems, as it allows for the integration of varied human experiences and values into the technology's design and application, leading to more equitable and socially attuned AI solutions.

Integrating a participatory approach with the precursors of trustworthy AI as outlined in the NIST AI Risk Management Framework (AI RMF) and the RFI fosters a comprehensive socio-technical approach. This integration is crucial in avoiding rapid technology development pathologies like market capture, monopolization, widespread distrust (but also overtrust), and lock-in. By including diverse societal inputs in the AI development process, we ensure that AI systems are not only technically robust and reliable but also socially and ethically aligned. This participatory dimension in AI development helps in preemptively addressing and mitigating issues that arise when technology development is solely market-driven or lacks diverse stakeholder involvement. Thus, a participatory approach is instrumental in creating AI systems that are not only effective and efficient but also equitable and representative of a broad range of societal needs and values.

VIII. Socio-technical Expertise at TRAILS

TRAILS researchers at the University of Maryland, George Washington University, Morgan State University, and Cornell University are using a participatory, socio-technical framework to identify and analyze AI trust challenges and opportunities and then to use the resulting insights to inform development of solutions. TRAILS research integrates both social-behavioral and technical approaches to understanding problems and designing solutions.

Analyzing AI Trust Challenges: Social and Behavioral

TRAILS researchers are using social and behavioral methodologies to shed light on the complexities of societal risks and impacts in specific domains by engaging stakeholders, investigating their actual experience with technology, individually and collaboratively and analyzing their social context. These inquires can inform both AI technology and policy development.

AI-Enabled Online Aggression and Children's Privacy

Virginia Byrne (Morgan State University) studies the trauma created by aggression in online spaces, and its effects on students. Exacerbated by the COVID pandemic, cyberbullying has become a significant source of trauma for children and youth (V. Byrne 2022). The use of generative AI to generate a large volume of harassing messages and to produce deepfakes, including hyper-realistic simulated pornographic images or videos of a targeted person, is likely to make the problem much worse. Because of a perceived risk of harm to perpetrators, victims are reluctant to report incidents to school officials or law enforcement, and turn to technical solutions, using features of social media platforms such as blocking and changing privacy settings (V. L. Byrne et al. 2023). However, reliance on these technical solutions creates unintended negative consequences, including self-silencing and conflict avoidance that leads to a loss of connection and social capital for young people who may already feel marginalized (V. L. Byrne 2021a, 2021b). Even just witnessing others being harassed online can lead children to shift towards defensive online engagement, including during remote learning activities (V. L. Byrne and Hollingsworth 2021). Byrne's research illustrates how purely technical solutions that fail to consider specific stakeholders' experience and their social and institutional contexts are unlikely to fully address critical safety concerns.[i] Designing better technical and social mitigation strategies will require a participatory approach, involving children directly in articulating their privacy needs (Kumar et al. 2023).

Socio-Technical Threats to the Information Ecosystem, from Manipulation to Polarization

Cody Buntain (University of Maryland) studies the online information space and the socio-technical problems and solutions that arise from algorithmic curation, content moderation, and how society relies on these spaces for collective sensemaking during moments of unrest and uncertainty. Beginning with studies of social media use during disaster, Buntain has studied how the society uses these spaces for collective coping (Buntain and Lim 2018) and technological solutions for making online spaces more informative during crises (C. Buntain et al. 2021;

McCreadie and Buntain 2022). Crucially, these and related studies have revealed substantial disconnects between technological capabilities and stakeholder needs (Lorini et al. 2021; Purohit, H. et al. 2024) and socio-technical problems around rumor propagation and the difficulty of corrections during these moments. Much of this work is applicable to online political engagement, especially information quality, resilience to manipulation, and the degree to which technologies impact social conditions. Buntain thus expanded into computational social science, studying how malevolent actors use online spaces to influence audiences (Alizadeh et al. 2020), how content moderation can have unexpected social consequences (C. Buntain et al. 2023), and how platform affordances interact with politicians' online behaviors (Buntain, C. et al. 2024). While online spaces have the potential to bring people together, provide social support, and accelerate collective understanding of the world, these spaces are falling well short of that potential, and the issues driving these shortcomings are socio-technical in nature and thus cannot be solved by technological solutions alone.

Use of models in disaster response and international development

Erica Gralla (George Washington University) has expertise and experience in studying the interaction between complex mathematical models and those who deploy them in practice. Her past work examines this interface between models and practice in the "extreme cases" of disaster response and international development: both contexts are ripe for the use of AI tools but place extreme value on trustworthiness (Blair et al. 2021; Gralla et al. 2014; Gralla and Goentzel 2018; Moline et al. 2019). These are therefore useful contexts in which to explore questions about AI-related risks and how to use both algorithmic and sociotechnical approaches to measure and mitigate them. In her past work, Gralla has used observational social science approaches to understand the human values and decision-making behavior that must be captured or accounted for in algorithmic models, then used that knowledge to drive the development of tools that fit these organizational, cognitive, and values-based constraints. Doing similar work on a larger scale, in collaboration with AI researchers, could help to develop essential new approaches for evaluating and managing AI risks.

Building Values-Driven AI that Supports Trustworthy and Accountable Decisions

Valerie F. Reyna (Cornell University) is an expert on decision-making involving risk and uncertainty (Reyna 2021; Reyna et al. 2023). Reyna has developed an evidence-based framework—fuzzy-trace theory—that has been applied to understand, predict, and evaluate a wide variety of high-stakes decisions and their implications for policy (Reyna 2023; Reyna et al. 2014). She has conducted extensive research on perceiving, mitigating, evaluating, and communicating risk in contexts involving technology, such as informatics in medicine and sharing of misinformation on AI-powered social-media platforms (Reyna 2023; Reyna et al. 2021). Her recent research concerns explaining and evaluating risks of AI and machine learning, especially socio-cognitive factors that promote trust, accountability, and values-driven decision-making (Edelson et al. 2023). This work provides a comprehensive and rigorous framework for understanding the social and behavioral foundations of explainability and interpretability of AI, which is directly relevant to NIST's interest in fostering values-driven AI systems that improve human lives, while building trust and accountability.

Social Dynamics of Explainability and Decision Making Informed by Psychological Foundations

David A. Broniatowski (George Washington University) conducts research in several areas pertinent to the NIST RFI. He conducts socio-technical evaluations of AI-powered social-media platforms' attempts to control misinformation spread, with his research highlighting the challenges of managing health misinformation in the digital space when audiences actively seek it out, and how these challenges differ for systems with different information flow architectures (Broniatowski et al. 2023). His insights into system architecture align with NIST's focus on developing systems that are responsive to the "ilities". Furthermore, understanding the influence of system design on user behavior and information flow is critical for AI system development. Including in collaboration with Reyna, Broniatowski also conducts work on the psychological foundations of explainability and interpretability in AI, addressing how these factors are key facilitators of trust in AI systems that are both social and technical in nature (Broniatowski 2021). This is directly relevant to NIST's interest in creating AI systems that are transparent and understandable to users. His work on how technical experts make decisions under risk provides valuable insights into how developers are likely to perceive complex engineered systems (Marti and Broniatowski 2020). This research is crucial for AI systems where risk assessment and management are integral. Beyond technical experts, his research on effectively communicating complex technical information to policymakers is essential for translating AI technical details into actionable insights for policy decisions (Broniatowski 2019). Finally, Broniatowski's research on assessing causal claims in complex engineered systems addresses the validity concerns in AI systems (Broniatowski and Tucker 2017). This work is significant for understanding the reliability and validity of AI evaluations. Collectively, these research contributions provide a comprehensive view of socio-technical considerations crucial for AI development and evaluation, aligning closely with the objectives of the NIST RFI. His work underscores the importance of considering psychological, social, and technical aspects in AI, addressing key NIST concerns about AI's broader impact.

Analyzing AI Trust Challenges: Technical

Informed by the context-rich examination of how people interact with AI within specific social contexts and the societal consequences, TRAILS researchers are also developing technical tools and methods for identifying and rigorously documenting safety issues, such as bias and robustness.

Detection and quantification of model bias

Despite increasing interest in studying the downstream impacts of model bias, fragility, and interpretability, technical tools for quantifying and controlling these model properties are still in their infancy. Building on the work of **Hal Daumé** (University of Maryland) and his team, TRAILS researchers are going beyond the superficial task of determining whether a model has "bias" to ask (1) what is the downstream economic or representational impact of model behavior, (2) what actions improve the model by mitigating impact, (3) under which range of circumstances

is it appropriate to use the model, and (4) what safeguards ensure the model is not deployed in an inappropriate context (Blodgett et al. 2020).

Unfair model behaviors often are driven by imbalances or biases in the datasets on which that model is trained; for example, the team led by **Jordan Boyd-Graber** (University of Maryland) showed that question answering datasets are overwhelmingly skewed toward American men (Gor et al. 2021), and Daume's team found that decades of coreference resolution work has a strong binary gender bias (Cao and Daumé III 2021). However, Goldstein's team found that disparities in model performance depend on both, and the relation between them (Cherepanova et al. 2022). This work demonstrates the importance of auditing model behaviors using datasets that reflect the population on which the model will be deployed.

Scalable Formal Verification for Correctness and Completeness

Research led by **Peng Wei** (George Washington University) research show that simulations and reasoning about models are valuable approaches to evaluating the safety of aviation systems that might also prove useful to analyzing the safety of generative AI. In order to scale verification for neural-network-embedded aviation systems, Wei is integrating adaptive stress testing (AST) into high fidelity simulators on selected use cases to accelerate falsification (Baheri et al. 2022). AST uses Gaussian process and reinforcement learning algorithms to efficiently sample the high-dimension scenario space for failure discovery. Peng's team is developing methods for inferring adequate component specifications from system-level requirements to enable compositional verification. In addition, they are exploring how to incorporate reasoning about the system (vs. treating it as a black box) in order to obtain stronger guarantees than claiming correctness on a finite set of simulations (Guo et al. 2022). For example, they are exploring techniques for proving robustness properties of neural networks by reasoning mathematically about their parameters rather than pointwise testing of input-output relationships. They will also infer barrier certificates or related artifacts which certify system correctness in a certain domain.

Designing solutions to socio-technical AI risks and opportunities: Social and Behavioral

Informed by analysis of risks and opportunities, TRAILS is using social and behavioral approaches to understand and design systems through which people and AI work together to develop solutions. They are also mapping out the emerging landscape of participatory AI projects across the AI ecosystem to identify promising methods for engaging stakeholders in developing new technology and policy.

Social and technical architectures of collaborative design and inclusion

Zoe Szajnfarber (George Washington University) is researching how the way you architect a technical or organizational system, and frame the problem, affects who can, and will, contribute novel solutions. This has implications for both participation and appropriate strategies for validating inputs from non-traditional sources. Past work has demonstrated that the different architectures better leverage contributions from experts, amateurs, specialists and human-AI

teams (Szajnfarber et al. 2022) and explored how the framing of design problems affects participation by underrepresented groups and organizations (Topcu et al. 2023; Vrolijk and Szajnfarber 2015). Current projects are characterizing risks and opportunities for incorporating LLMs into the design workflow (Szajnfarber, Zoe et al. 2023), and examining how interpretability by non-AI experts affects which AI enabled systems are being selected in the DoD acquisition process (Krueger, Chris et al. 2023).

Value-Centered Design and Participatory AI Approaches

Katie Shilton (University of Maryland) researches value-centered design (Shilton 2018), a participatory approach which requires close collaboration with stakeholders and shared responsibility for ethical concerns such of power, inequity, and trust (Shilton and Anderson 2017). Shilton has shown the efficacy of this approach for curation of training data for AI models intended to assess cultural and interpersonal sensitivity (Iqbal et al. 2021). Through also investigating the range of scope of participatory AI projects and initiative around the world, Shilton shows consistent use of qualitative methods, such as workshops, interviews, and agile design and prototyping (Arango and Shilton 2024). However, participatory AI in the global south more frequently adopt critical "lenses" focused on addressing power imbalances in how AI is designed and applied. In contrast, those in the global north are more likely to foreground usability and effectiveness of the technology divorced from consideration of larger structural inequities. Capable of incorporating many common methods used across projects, ethnographic methodology is particularly promising for future participatory AI work because it requires critically examines issues of power (Shilton et al. 2021).

Participatory AI Policy Development

Susan Aaronson (George Washington University) is researching ways and the extent to which governments are actively engaging the public in developing AI policy to increase safety. This work suggests that policy makers globally are not yet taking a systematic approach to resolving key policy issues raised by new developments in generative AI, including the IP and privacy concerns related to web scraping in model development and the implications of open versus closed source approach to the quality and validity of datasets (Aaronson 2023). She and her team are also developing recommendations on making policy making more participatory, better reflecting the sociotechnical realities of how these technologies affect citizens. For example, Aaronson and Zable examined whether policymakers consulted with and "heard" public comment on AI strategies. They found most countries with an AI strategy consulted, but only four actually included those comments, and none changed their strategy in response to public comments (Aaronson and Zable 2023). To be truly participatory, policymakers would have to show they really listen. A socio-technical approach grounded in truly deliberative democratic practices would ensure that policymakers educate and involve their citizenry, creating a trusted feedback loop.

Designing solutions to socio-technical AI risks and opportunities: Technical

Although a technical approach is not sufficient to fully address AI safety, sophisticated technical innovations are a necessary component of comprehensive solutions. TRAILS researchers are developing methods for detecting synthetic content and enabling more effective and equitable integration of human and AI capabilities.

Detecting Synthetic Content

Identifying synthetic content: Fake and manipulated images

Research led by **Abhinav Shrivastava** (University of Maryland) responds to recent advances that have democratized access to sophisticated tools for creating fake images, videos, audio, and text. While there are many exciting applications of these technologies, their use in creating large-scale disinformation campaigns that often elicit emotional responses and sow discord, poses significant threats. Even though automated detection of falsified, manipulated, and deceptive information is the first-line defense in the fight against disinformation campaigns, traditional approaches have focused on developing piecemeal statistical tools that can be utilized by experts (e.g., journalists) or organizations (e.g., social media firms, independent fact-checking firms). For example, Shrivastava has worked on developing tools to detect fake and manipulated images and videos and collaborates with industry partners to deploy these tools in practice. For example, the technology incubator Jigsaw (Google Inc.) developed 'Assembler' to equip journalists with sophisticated fact-checking tools, which included an algorithm by Shrivastava for detecting image manipulation (Zhou et al. 2019). However, it is generally the responsibility of experts to piece together disparate evidence to support/refute information, which is time-consuming. Similarly, Shrivastava collaborated with Meta Inc. (formerly Facebook Inc.) to deploy these tools internally to detect manipulated media for critical elections in India and EU (Wang et al. 2022; Zhou et al. 2019, 2021). These tools are often deployed internally to prevent the detected media from propagating the social network. A critical component missing is the use of automated tools for flagging/rating trustworthiness of content for end-users who consume media. Research in this requires a socio-technical approach to adapt these tools to become indispensable for end-users when consuming media on their devices. The goal of this socio-technical approach is to educate netizens about the impact of AI, and to preserve and return their trust to digital media.

Identifying synthetic content: Text

Furong Huang (University of Maryland) researches methods for distinguishing between human-written text and outputs from Large Language Models (LLMs). Detection, albeit difficult, is indeed possible. This research demonstrates that with substantial text data, either through increased sample sizes or extended sequence lengths, the feasibility of distinguishing between human and AI-generated text improves (Chakraborty et al. 2023). This insight, gained through rigorous empirical testing across various datasets and state-of-the-art text generators, is a significant stride in the field. However, the inherent difficulty of this task often necessitates more robust solutions, such as watermarking, to enhance reliability in detection.

Watermarking synthetic text

Tom Goldstein (University of Maryland) has developed an innovative watermarking framework that embeds undetectable signals into the text generated by proprietary LLMs (Kirchenbauer et al. 2023). These watermarks are imperceptible to humans but can be algorithmically identified, offering a novel approach to secure and authenticate AI-generated content. Tested on models like the Open Pretrained Transformer (OPT) family, this method demonstrates a minimal impact on text quality while ensuring efficient and reliable detection.

However, an important consideration in the realm of watermarking is its vulnerability to adversarial attacks. For example, research by **Soheil Feizi** (University of Maryland) shows that recursive paraphrasing can be effective against watermarking, as well as other methods for detecting synthetic text (Sadasivan et al. 2023). This susceptibility underscores the necessity of rigorously stress-testing watermarks to evaluate and enhance their robustness.

Huang and **Goldstein** have collaborated on research responds to this need by developing WAVES, a comprehensive toolkit designed to assess the resilience of watermarking systems under a variety of challenging scenarios (An et al. 2024). WAVES subjects watermarks to an array of attacks, ranging from traditional image distortions to more advanced adversarial methods. This stress-testing not only reveals previously undetected vulnerabilities in modern watermarking algorithms but also aids in the development of more robust watermarking techniques. By providing a diverse set of stress tests and integrating them into a standardized evaluation protocol, WAVES aims to benchmark and improve the resilience of watermarks, ensuring their effectiveness in the face of sophisticated digital threats.

Technical Approaches to enabling Trustworthy Human-AI Collaboration

Human-AI Collaboration in Model Development, Application and Evaluation

The team led by **Hal Daumé** (University of Maryland) is developing technical tools and techniques for human-AI collaboration that both improve performance and increase equity. Their research is intended to help people understand and evaluate the information and recommendations AI provides. The weight of evidence method uses incremental explanations to explain complex decisions (Alvarez Melis et al. 2021), providing contrastive explanations—why answer is both true and false—helps people make more accurate evaluations of the output of LLMs (Si et al. 2023), and automated presentation of relevant background information used by the model helps users better evaluate the predictions of AI QA systems (Goyal et al. 2023). Daumé's team's research analyzes the suitability of AI models to the challenges faced by people working with them to solve real-world problems, such as applying community-generated rules in AI-assisted content moderation (Cao et al. 2023), in order to inform development of tools that address their needs, are aligned with their sense making processes, and are likely to increase equity. In addition to creating tools and techniques that help people use AI more effectively, the team has also developed ways for humans to help AI improve its own performance, such as through enabling reinforcement-learning-based methods to model when

they are no longer able to make progress and request help from users (Nguyen et al. 2022; Nguyen and Daumé III 2019).

References

Aaronson, S. A. (2023, August 28). Data Dysphoria: The Governance Challenge Posed by Large Learning Models. SSRN Scholarly Paper, Rochester, NY. https://doi.org/10.2139/ssrn.4554580

Aaronson, S. A., & Zable, A. (2023, August 28). Missing Persons: The Case of National AI Strategies. SSRN Scholarly Paper, Rochester, NY. https://doi.org/10.2139/ssrn.4554650

Alizadeh, M., Shapiro, J. N., Buntain, C., & Tucker, J. A. (2020). Content-based features predict social media influence operations. *Science Advances*, *6*(30), eabb5824. https://doi.org/10.1126/sciadv.abb5824

Alvarez Melis, D., Kaur, H., Daumé Iii, H., Wallach, H., & Wortman Vaughan, J. (2021). From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *9*, 35–47. https://doi.org/10.1609/hcomp.v9i1.18938

An, B., Ding, M., Rabbani, T., Agrawal, A., Xu, Y., Deng, C., et al. (2024, January 22). Benchmarking the Robustness of Image Watermarks. arXiv. https://doi.org/10.48550/arXiv.2401.08573

Baheri, A., Ren, H., Johnson, B., Razzaghi, P., & Wei, P. (2022, May 14). A Verification Framework for Certifying Learning-Based Safety-Critical Aviation Systems. arXiv. https://doi.org/10.48550/arXiv.2205.04590

Blair, C., Gralla, E., Wetmore, F., Goentzel, J., & Peters, M. (2021). A Systems Framework for International Development: The Data-Layered Causal Loop Diagram. *Production and Operations Management*, *30*(12), 4374–4395. https://doi.org/10.1111/poms.13492

Blodgett, S. L., Barocas, S., Daumé Iii, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Broniatowski, D. A. (2019). Communicating Meaning in the Intelligence Enterprise. *Policy Insights from the Behavioral and Brain Sciences*, *6*(1), 38–46. https://doi.org/10.1177/2372732218792061

Broniatowski, D. A. (2021). Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. *NIST*. https://www.nist.gov/publications/psychological-foundations-explainability-and-interpretability-artificial-intelligence. Accessed 26 January 2024

Broniatowski, D. A., Simons, J. R., Gu, J., Jamison, A. M., & Abroms, L. C. (2023). The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic. *Science Advances*, *9*(37), eadh2132. https://doi.org/10.1126/sciadv.adh2132

Broniatowski, D. A., & Tucker, C. (2017). Assessing causal claims about complex engineered systems with quantitative data: internal, external, and construct validity. *Systems Engineering*, *20*(6), 483–496. https://doi.org/10.1002/sys.21414

Buntain, C., Greene, K., DeVerna, M., & Tucker, J. (2024). *Hot Tweets and Cold Posts: Politicians' Ideological Presentations on Twitter and Facebook* (Tech Report).

Buntain, C., Innes, M., Mitts, T., & Shapiro, J. (2023). Cross-Platform Reactions to the Post-January 6 Deplatforming. *Journal of Quantitative Description: Digital Media*, *3*. https://doi.org/10.51685/jqd.2023.004

Buntain, C. L., & Lim, J. K. R. (2018). #pray4victims: Consistencies in Response to Disaster on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 25:1-25:18. https://doi.org/10.1145/3274294

Buntain, C., McCreadie, R., & Soboroff, I. (2021). Incident Streams 2021 off the Deep End: Deeper Annotations and Evaluations in Twitter.

Byrne, V. (2022). Introduction to the Special Issue: Critical Perspective on Online Trauma. *Journal of Trauma Studies in Education*, *1*(2), 1–3.

Byrne, V. L. (2021a). "You might as well just all agree with each other:"An initial study of cyberbullying victims' social presence in online discussions. *Computers & Education*, *167*, 104174. https://doi.org/10.1016/j.compedu.2021.104174

Byrne, V. L. (2021b). Blocking and Self-Silencing: Undergraduate Students' Cyberbullying Victimization and Coping Strategies. *TechTrends*, *65*(2), 164–173. https://doi.org/10.1007/s11528-020-00560-x

Byrne, V. L., & Hollingsworth, J. (2021). An Initial Empirical Study of Witnessing Online Harassment and Experiencing Secondary Trauma Among College Students. *Technology and Higher Education*, *2*(1), 19–32.

Byrne, V. L., Hollingsworth, J., & Kumar, P. C. (2023). Navigating privacy tensions when responding to online aggression at postsecondary institutions. *British Journal of Educational Technology*, *54*(6), 1636–1652. https://doi.org/10.1111/bjet.13377

Cao, Y. T., & Daumé III, H. (2021). Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics*, *47*(3), 615–661. https://doi.org/10.1162/coli_a_00413

Cao, Y. T., Domingo, L.-F., Gilbert, S. A., Mazurek, M., Shilton, K., & Daumé III, H. (2023, November 13). Toxicity Detection is NOT all you Need: Measuring the Gaps to Supporting Volunteer Content Moderators. arXiv. https://doi.org/10.48550/arXiv.2311.07879

Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., & Huang, F. (2023, October 2). On the Possibilities of AI-Generated Text Detection. arXiv. https://doi.org/10.48550/arXiv.2304.04736

Cherepanova, V., Reich, S., Dooley, S., Souri, H., Goldblum, M., & Goldstein, T. (2022, March 15). A Deep Dive into Dataset Imbalance and Bias in Face Identification. arXiv. https://doi.org/10.48550/arXiv.2203.08235

Edelson, S. M., Roue, J. E., Singh, A., & Reyna, V. F. (2023). How Decision Making Develops: Adolescents, Irrational Adults, and Should AI be Trusted With the Car Keys? *Policy Insights from the Behavioral and Brain Sciences*, 23727322231220423. https://doi.org/10.1177/23727322231220423

Farke, F. M., Balash, D. G., Golla, M., Dürmuth, M., & Aviv, A. J. (2021). Are Privacy Dashboards Good for End Users? Evaluating User Perceptions and Reactions to Google's My Activity (pp. 483–500). Presented at the 30th USENIX Security Symposium (USENIX Security 21). https://www.usenix.org/conference/usenixsecurity21/presentation/farke. Accessed 26 January 2024

Gary A Marchant & Wendell Wallach. (2015, Summer). Coordinating Technology Governance. *Issues in Science and Technology*, *XXXI*(4). https://issues.org/coordinating-technology-governance/. Accessed 2 February 2024

Gor, M., Webster, K., & Boyd-Graber, J. (2021). Toward Deconfounding the Effect of Entity Demographics for Question Answering Accuracy. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 5457–5473). Presented at the EMNLP 2021, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.444

Goyal, N., Briakou, E., Liu, A., Baumler, C., Bonial, C., Micher, J., et al. (2023, October 25). What Else Do I Need to Know? The Effect of Background Information on Users' Reliance on QA Systems. arXiv. http://arxiv.org/abs/2305.14331. Accessed 1 February 2024

Gralla, E., & Goentzel, J. (2018). Humanitarian transportation planning: Evaluation of practice-based heuristics and recommendations for improvement. *European Journal of Operational Research*, *269*(2), 436–450. https://doi.org/10.1016/j.ejor.2018.02.012

Gralla, E., Goentzel, J., & Fine, C. (2014). Assessing Trade-offs among Multiple Objectives for Humanitarian Aid Delivery Using Expert Preferences. *Production and Operations Management*, *23*(6), 978–989. https://doi.org/10.1111/poms.12110

Guo, W., Zhou, Y., & Wei, P. (2022). Exploring online and offline explainability in deep reinforcement learning for aircraft separation assurance. *Frontiers in Aerospace Engineering*, *1*. https://www.frontiersin.org/articles/10.3389/fpace.2022.1071793. Accessed 26 January 2024

Iqbal, M., Shilton, K., Sayed, M. F., Oard, D., Rivera, J. L., & Cox, W. (2021). Search with Discretion: Value Sensitive Design of Training Data for Information Retrieval. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 133:1-133:20. https://doi.org/10.1145/3449207

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A Watermark for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 17061–17084). Presented at the International Conference on Machine Learning, PMLR. https://proceedings.mlr.press/v202/kirchenbauer23a.html. Accessed 26 January 2024

Krueger, Chris, Manning, Justin, Pless, Robert, & Szajnfarber, Zoe. (2023). All Models Fail, But Some Are Useful: Enabling Informed Decision-Making by Non-Expert Acquirers of AI-Embedded Systems. Presented at the CESUN 2023: 9th International Engineering Systems Symposium, Evanston, IL.

Kumar, P. C., O'Connell, F., Li, L., Byrne, V. L., Chetty, M., Clegg, T. L., & Vitak, J. (2023). Understanding Research Related to Designing for Children's Privacy and Security: A Document Analysis. In *Proceedings of the 22nd Annual ACM Interaction Design and*

Children Conference (pp. 335–354). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3585088.3589375

Lorini, V., Castillo, C., Peterson, S., Rufolo, P., & Purohit, H. (2021). Social Media for Emergency Management: Opportunities and Challenges at the Intersection of Research and Practice.

Maria Isabel Magana Arango & Katie Shilton. (2024). In It Together? A Systemic Analysis of Particitory AI in Global Projects. Presented at the AI-HCI.

Marti, D., & Broniatowski, D. A. (2020). Does gist drive NASA experts' design decisions? *Systems Engineering*, *23*(4), 460–479. https://doi.org/10.1002/sys.21538

Mathews, D. J. H., Fabi, R., & Ii, A. C. O. (2022). Imagining Governance for Emerging Technologies. *Issues in Science and Technology*, *XXXVIII*(3). https://issues.org/imagining-governance-emerging-technologies-mathews-fabi-offodile/

McCreadie, R., & Buntain, C. (2022). CrisisFACTS: Buidling and Evaluating Crisis Timelines.

Moline, J., Goentzel, J., & Gralla, E. (2019). Approaches for Locating and Staffing FEMA's Disaster Recovery Centers. *Decision Sciences*, *50*(5), 917–947. https://doi.org/10.1111/deci.12359

Nguyen, K., Bisk, Y., & Daumé III, H. (2022, June 22). A Framework for Learning to Request Rich and Contextually Useful Information from Humans. arXiv. https://doi.org/10.48550/arXiv.2110.08258

Nguyen, K., & Daumé III, H. (2019, November 22). Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning. arXiv. https://doi.org/10.48550/arXiv.1909.01871

Purohit, H., Buntain, C., Hughes, A., Peterson, S., Lorini, V., & Castillo, C. (2024). Engage and Mobilize! How Emergency Management Use Social Media. In *Submitted to CSCW*.

Reyna, V. F. (2021). A scientific theory of gist communication and misinformation resistance, with implications for health, education, and policy. *Proceedings of the National Academy of Sciences*, *118*(15), e1912441117. https://doi.org/10.1073/pnas.1912441117

Reyna, V. F. (2023). Social media: Why sharing interferes with telling true from false. *Science Advances*, *9*(9), eadg8333. https://doi.org/10.1126/sciadv.adg8333

Reyna, V. F., Broniatowski, D. A., & Edelson, S. M. (2021). Viruses, Vaccines, and COVID-19: Explaining and Improving Risky Decision-making. *Journal of Applied Research in Memory and Cognition*, *10*(4), 491–509. https://doi.org/10.1016/j.jarmac.2021.08.004

Reyna, V. F., Chick, C. F., Corbin, J. C., & Hsia, A. N. (2014). Developmental Reversals in Risky Decision Making: Intelligence Agents Show Larger Decision Biases Than College Students. *Psychological Science*, *25*(1), 76–84. https://doi.org/10.1177/0956797613497022

Reyna, V. F., Müller, S. M., & Edelson, S. M. (2023). Critical tests of fuzzy trace theory in brain and behavior: uncertainty across time, probability, and development. *Cognitive, Affective, & Behavioral Neuroscience*, *23*(3), 746–772. https://doi.org/10.3758/s13415-022-01058-0

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023, June 28). Can AI-Generated Text be Reliably Detected? arXiv. https://doi.org/10.48550/arXiv.2303.11156

Shilton, K. (2018). Values and Ethics in Human-Computer Interaction. *Foundations and Trends® in Human–Computer Interaction*, *12*(2), 107–171. https://doi.org/10.1561/1100000073

Shilton, K., & Anderson, S. (2017). Blended, Not Bossy: Ethics Roles, Responsibilities and Expertise in Design. *Interacting with Computers*, *29*(1), 71–79. https://doi.org/10.1093/iwc/iww002

Shilton, K., Moss, Emanuel, Gilber, Sarah A., Bietz, Matthew J., Metcalf, Jacob, Vitak, Jessica, & Zimmer, Michael. (2021). Excavating awareness and power in data science: A manifesto for trustworthy pervasive data research - Katie Shilton, Emanuel Moss, Sarah A. Gilbert, Matthew J. Bietz, Casey Fiesler, Jacob Metcalf, Jessica Vitak, Michael Zimmer, 2021. *Big Data and Society*, (July-December), 1–12.

Si, C., Goyal, N., Wu, S. T., Zhao, C., Feng, S., Daumé III, H., & Boyd-Graber, J. (2023, October 19). Large Language Models Help Humans Verify Truthfulness -- Except When They Are Convincingly Wrong. arXiv. http://arxiv.org/abs/2310.12558. Accessed 1 February 2024

Szajnfarber, Z., Topcu, T. G., & Lifshitz-Assaf, H. (2022). Towards a solver-aware systems architecting framework: leveraging experts, specialists and the crowd to design innovative complex systems. *Design Science*, *8*, e10. https://doi.org/10.1017/dsj.2022.7

Szajnfarber, Zoe, Adeyeye, Olandele, & Pless, Robert. (2023). Opportunities and Risks of Incorporating LLMs in the Systems Engineering and Design Workflow: A Case Study of Robotic System Design Process. Presented at the AI4SE & SE4AI Workshop 2023.

Topcu, T. G., Zhang, L. "Lydia," & Szajnfarber, Z. (2023). Does Open Innovation Open Doors for Underrepresented Groups to Contribute to Technology Innovation?: Evidence from a Space Robotics Challenge. *Space Policy*, *64*, 101550. https://doi.org/10.1016/j.spacepol.2023.101550

Vrolijk, A., & Szajnfarber, Z. (2015). When Policy Structures Technology: Balancing upfront decomposition and in-process coordination in Europe's decentralized space technology ecosystem. *Acta Astronautica*, *106*, 33–46. https://doi.org/10.1016/j.actaastro.2014.10.017

Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S.-N., & Jiang, Y.-G. (2022, March 29). ObjectFormer for Image Manipulation Detection and Localization. arXiv. https://doi.org/10.48550/arXiv.2203.14681

Zhou, P., Chen, B.-C., Han, X., Najibi, M., Shrivastava, A., Lim, S. N., & Davis, L. S. (2019, September 11). Generate, Segment and Refine: Towards Generic Manipulation Segmentation. arXiv. https://doi.org/10.48550/arXiv.1811.09729

Zhou, P., Yu, N., Wu, Z., Davis, L. S., Shrivastava, A., & Lim, S.-N. (2021, January 26). Deep Video Inpainting Detection. arXiv. https://doi.org/10.48550/arXiv.2101.11080

[i] Although in this case use of privacy features create unintended consequences, another danger of such features when designed without the participation of end users and without understanding of their social context is that they won't be used at all, as Adam Aviv (George Washington University) and his collaborators show is the case on one popular platform (Farke et al. 2021). The inclusion of features so designed makes it look like companies are improving privacy when they are not in actual practice.