Problematic AI: Finding the best way forward

Shenefiel, Chris
CoVA CCI Senior Cyber Law Researcher, Adjunct Lecturer
cashenefiel@wm.edu

Burkley, Sally
J.D. Candidate (2025)
sburkley@wm.edu

Collins, Kathleen
J.D. Candidate (2025)
kscollins02@wm.edu

Shin, Daniel
Cybersecurity Researcher, CoVA CCI Research Scientist
dshin01@wm.edu

William & Mary Law School, Center for Legal and Court Technology

July 7, 2023

Contents

1 Background

Beginning with his seminal 1955 summer research project on Artificial Intelligence, John McCarthy at Dartmouth College, started a revolution in Computer Science, technology and, ultimately, society.[1] This bold Artificial Intelligence research project began when computers were very new and only just starting to perform basic capabilities like time-sharing and multi-tasking. At the Dartmouth College AI@50 conference, James Moore described the impact of the first conference and the evolution of AI over the years in learning, vision, language, game playing and robotics.[2] In the intervening 68 years, AI has made amazing progress. For instance, AI has been shown to be able to look through walls to determine the position, posture and movement of people[3]. AI research has been able to establish multi-person brain-to-brain interfaces for collaboration,[4] can lip-read[5] and detect hidden emotions through facial micro-expressions.[6] However, with all these remarkable advancements, the unique nature of AI and Machine Learning can still result in unexpected outcomes.

Unexpected results are, unfortunately, a natural aspect of the way AI operates. In a 2020 paper, Hagendorff and Wezel summarized 15 challenges for AI.[7] The challenges center around how models are trained, insufficient understanding of the risks associated with employing AI, the limited expandability of AI systems, and model fragility. But why is AI so different from classical programming and why does it suffer from these unique challenges?

The computer science paradigm of Artificial Intelligence is substantially different from classical programming. Classical programming (e.g., programming an application with prescribed logic instructions and data) is *deterministic* because the program will only perform according to the logic written in the code. While not always correct or error-free, its behavior is predictable based on the expressed intent of the programmer represented in the code. Artificial Intelligence-based systems employ *probabilistic algorithms* using symbolic versus quantitative programming. These AI algorithms adjust probabilities (simulated neural connections in Neural Networks) based on the training data. During supervised learning, a human engineer labels the training data so that the model learns to match a specific pattern with a label.[8] For instance, a model designed to detect highway speed limit signs, would be trained in millions of variations of traffic sign images, each properly labeled. The engineer is not creating code that describes the characteristics of a street sign. Instead, the inference model is developing internal patterns based on the different sets of inputs it receives during training. When a model successfully classifies an object, we don't always know how

[1] John McCarthy. "A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE". in: (1955). URL: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

[2] James Moor. "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years | AI Magazine". In: (2006). URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1911.

[3] Jiancun Zuo et al. "A New Method of Posture Recognition Based on WiFi Signal". In: *IEEE Communications Letters* 25.8 (2021), pp. 2564–2568. DOI: 10.1109/LCOMM.2021.3081135.

[4] Linxing Jiang et al. "BrainNet: A Multi-Person Brain-to-Brain Interface for Direct Collaboration Between Brains". In: *Scientific Reports* 9.1 (Apr. 2019). DOI: 10.1038/s41598-019-41895-7. URL: https://doi.org/10.1038%2Fs41598-019-41895-7.

[5] Brendan Shillingford et al. *Large-Scale Visual Speech Recognition*. 2018. arXiv: 1807.05162 [cs.CV].

[6] Xiaobai Li et al. "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods". In: *IEEE transactions on affective computing* 9.4 (2017), pp. 563–577.

[7] Thilo Hagendorff and Katharina Wezel. "15 challenges for AI: or what AI (currently) can't do". In: *AI & SOCIETY* 35.2 (2020), pp. 355–365. ISSN: 1435-5655. DOI: 10.1007/s00146-019-00886-y. URL: https://doi.org/10.1007/s00146-019-00886-y.

[8] M. Dudziak, T. Fields, and K. Youngmann. "AI programming vs. conventional programming for autonomous vehicles - Trade-off issues". In: *Proceedings of the 1985 4th International Symposium on Unmanned Untethered Submersible Technology*. Vol. 4. 1985, pp. 284–296. DOI: 10.1109/UUST.1985.1158553.

it came to its conclusion because these internal patterns can be quite complex. Consequently, AI models can often lack transparency or explainability.[9]

Even if an AI model were producing correct answers it may not reflect reality. For instance, researchers from the University of Michigan found that slight distortions of stop signs (e.g., applying tape or shadings) could cause a highly accurate image detection model for self-driving cars to read a stop sign as a 45 mph speed limit sign.[10] A human seeing this perturbed stop sign would never interpret it this way, but the model, after developing its own set of characteristics for detecting a stop sign, was easily fooled with simple distortions. Therefore, models that may behave accurately can still yield unexpected and potentially dangerous results.

Given these challenges, and the various perspectives surrounding the development and use of AI systems, the Center for Legal & Court Technology (CLCT) at William & Mary Law School embarked on a conference to bring together technologists, policy-makers, social scientists, and lawyers to explore these limitations and consider the best direction forward.

2 Introduction

On February 10th, William & Mary Law School's (CLCT) offered a hybrid symposium entitled "Problematic AI: Finding the best way forward." The conference originated from an article by Professor Fred Lederer, "*Problematic AI – When Should We Use It*".[11] The symposium was graciously sponsored by COVA CCI [12] and the University of Montreal's CyberJustice Laboratory. [13] The conference discussed the degree to which fallible human beings should rely on fallible AI systems.

Please note that planning for this conference occurred before ChatGPT release. Therefore, we intentionally chose to focus on predictive AI. Generative AI may be the key topic for next year's conference.

[9] ES Vorm and David JY Combs. "Integrating Transparency, Trust, and Acceptance: The Intelligent Systems Technology Acceptance Model (ISTAM)". in: *International Journal of Human–Computer Interaction* 38.18-20 (2022), pp. 1828–1845.

[10] Kevin Eykholt et al. "Robust physical-world attacks on deep learning visual classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1625–1634.

[11] Fredric Lederer. "Problematic AI – When Should We Use It? — ALI Social Impact Review". In: (2023). URL: `https://www.sir.advancedleadership.harvard.edu/articles/problematic-ai-when-should-we-use-it`.

[12] Commonwealth Cybersecurity Initiative Coastal Virginia: `https://covacci.org/`

[13] Cyberjustice Laboratory: `https://www.cyberjustice.ca/`

3 Conference Summary

Panelists representing law, government, academia and industry from the United States, Canada and the United Kingdom participated in three panels to offer their respective insights on following three questions:

Panel 1: What is AI, how is it used and how can it produce erroneous results?	
Dr. Laura Freeman	Director, Intelligent Systems Division – National Security Institute; Research Associate Professor, Statistics Department, Virginia Tech
Dr. Blake Anderson	Principal Engineer - AI, Cisco Systems, Inc.
Greg Akers	Owner, Greg Akers Consulting
Dr. Emre Kazim	Chief Operating Officer, Holistic AI
Dr. Yan Lu (Panel Moderator)	Research Assistant Professor, School of Cybersecurity, Old Dominion University

Panel 2: How to detect and manage AI errors and risks; human baseline and engagement?	
Dennis Hirsch, J.D.	Director, Program on Data and Governance; Professor of Law, Moritz College of Law, The Ohio State University
Nicolas Vermeys, J.D.	Director of the Cyberjustice Laboratory; Full Professor, Université de Montréal
Katie Shay, J.D.	Associate General Counsel & Director of Human Rights, Cisco Systems, Inc.
Steven Truitt	Principal Program Manager AI, Microsoft
Dr. Abby Gilbert	Institute for the Future of Work (United Kingdom)
Dr. Iria Giuffrida (Moderator)	Professor of the Practice of Law, William & Mary Law School

Panel 3: How well do current regulations/policies address these challenges?	
Jessica Smith	Technology Policy Manager, Ofcom (United Kingdom)
Reva Schwartz	Research Scientist & Principal Investigator for AI Bias at National Institute of Standards and Technology (NIST)
Brenda Leong, J.D.	Partner, BNH.AI
Peter Chapman, J.D.	Associate Director and Tech and Human Rights Lead, Article One Advisors
Dr. Stephanie J. Blackmon (Moderator)	Associate Professor of Higher Education, William & Mary School of Education

Workshop: Best path forward
The final session of the day was a workshop for the remaining panelists and conference attendees to discuss the best path forward and next steps.

3.1 Panel 1: What is AI, how is it used and how can it produce erroneous results?

The first panel consisted of the following members of the Artificial Intelligence (AI) industry: Greg Akers (Executive Technology Consultant and Former Cisco SVP, Advanced Security Research/Governments), Dr. Blake Anderson (Cisco Principal Security Research Engineer), Dr. Laura Freeman (Director, Intelligent Systems Division – National Security Institute; Research Associate Professor, Statistics Department), and Dr. Emre Kazim (Holistic AI[14] Co-CEO and Co-Founder).

Mr. Akers mentioned that, despite media hype, AI is rarely solely relied upon for decision-making. Due to high-risk situations and financial stakes, AI use is generally limited to low-stakes areas or used as a tool that supports (not supplants) decision-making. AI is changing the way people work and get to decisions, but in high-stakes situations, such as in the military or legal context, a final human decision-maker is required and is often present in collaboration with AI. According to Akers, the key area of fault within the current scheme of AI is the lack of security; more important than ever because of code/data complexity and adversarial AI.[15] AI models are particularly vulnerable to over-training (resulting in incorrect answers), or data poisoning (when a competitor or adversary inserts incorrect data during model training).

Continuing Mr. Akers' discussion on data poisoning, Dr. Anderson emphasized the problems created from the training stage due to mislabeling and tainted data. For example, in cybersecurity, distinguishing malware signals from benign ones is challenging when there is label noise (e.g., where the system misidentifies benign signals as malware and vice versa).[16] Anderson stated that when noise levels are low, the problem is manageable. However, he pointed out that the problem becomes more unmanageable as noise levels increase. This is exacerbated when competitors or adversaries strategically insert noise, making the problem even more difficult to resolve.

Dr. Freeman focused on "who is responsible" for model errors. She explained that models can be brittle (e.g., identifying a pig as an airplane) because we don't always know what features the model is using. To help solve the problem of model errors, Dr. Freeman proposed to identify the important data features that the model should use so that we can guide the model to learn in the way we want. Furthermore, she highlighted that AI does not act independently but instead is dependent on interactions with humans, other systems, and operational situations/environments. In the realm of AI and automated systems, our national priorities with these technologies should be responsible, equitable, traceable, and reliable systems. She also emphasized the importance of training adequacy, data and model ownership, and managing the learning process. She pointed out that incorrect detection should be easily mitigated through testing and evaluating before AI system deployment rather than dealing with the ramifications afterward. Lastly, she cautioned against deploying AI in high risk circumstances but rather using it as a tool to help human operators to digest large amounts of data to speed decision-making.

Dr. Kazim was particularly interested in the problem of trustworthy AI. How can the engineering community trust AI? How can the law and society trust AI to be explainable and fair? "How do we do meaningful engineering assessments but with an eye to non-engineers?" To explore these questions, Dr. Kazim presented the current state of AI risk management. For example, he highlighted how some industries and government departments are engaged in the certification of

[14] Holistic AI: `https://www.holisticai.com/`

[15] Teng Long et al. "A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions". In: *Computers & Security* (2022), p. 102847.

[16] Justin M Johnson and Taghi M Khoshgoftaar. "A survey on classifying big data with label noise". In: *ACM Journal of Data and Information Quality* 14.4 (2022), pp. 1–43.

service for algorithms, such as the London Stock Exchange and England's Treasury, even without the certification standards. He stated that it is difficult to assess data and algorithms without standards. Unfortunately, assessments are needed nonetheless. Dr. Kazim suggested that one way to get to a baseline standard is through synthetic datasets that can be used to test algorithms and perform algorithm auditing. These techniques could help clarify how and when to regulate algorithm use. He also noted there is a delicate tension between model transparency and security. The more transparent a model, the more likely it may be exploited. Lastly, Dr. Kazim emphasized that any standards that are designed to prevent potential AI harm must not stifle AI benefit and innovation.

3.2 Panel 2: How to detect and manage AI errors and risks; human baseline and engagement?

In this panel, speakers from business and academia discussed possible ways to manage risks associated with AI, through both government regulation and internal management. Steven Truitt (Principal Program Manager for Microsoft) noted that, oftentimes, the "problematic" aspect of AI is not the technology *per se*, but the fact that it's being used in problematic ways or for problematic purposes. Thus, when we think about problematic AI, we need to consider the context in which the technology operates and how it interacts with human beings. The problem can be the system's misuse – either by using the technology for nefarious purposes or by users, who are unaware of the system's limitations, applying the technology inappropriately. For example, he explained the strengths and limitations of Large Language Model (LLM) generative AI chatbots (e.g., ChatGPT[17]). These systems are designed to provide interactive resources for pro/con arguments, hypothetical reasoning, summarization, story generation, personalization, simplification and clarification, and education. However, they are not designed to be precise and accurate sources for recent event information or for citations. When LLMs produce text that is inaccurate or apparently contrived, they are said to be "hallucinating".[18] Misuse of AI will very likely result in unsatisfactory results.

Dr. Abigail Gilbert (Head of Research for the Institute for the Future of Work)[19] noted that another area where the industry has seen AI misuse is in the workplace. While technology is expected to improve productivity and remove unpleasant or dangerous work, rather, AI can be used to drive the workforce to work harder in order to keep up.[20] To mitigate these harms, Dr. Gilbert emphasized the importance of a human-centered approach to the AI ecosystem; ensuring that this technology works for the public interest. She recommended creating a corporate duty to undertake Algorithmic Impact Assessments (AIAs) designed to forecast how the AI might impact vulnerable groups and to continuously monitor and evaluate the impact of the technology on workers.[21]

This AIA process would be undertaken in four stages: first, identifying and involving relevant stakeholders who should be engaged in determining what the impacts of a system could be; second, undertaking an *ex-ante* risk assessment; third, developing mitigations for those risks; and fourth, continuously monitoring and evaluating the impacts of AI systems. Most importantly, Dr. Gilbert

[17] OpenAI. "Introducing ChatGPT". in: (2023). URL: `https://openai.com/blog/chatgpt`.

[18] Yejin Bang et al. "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity". In: *arXiv preprint arXiv:2302.04023* (2023).

[19] Institute for the Future of Work: `https://www.ifow.org/`

[20] A Gilbert and A Thomas. "The Amazonian Era–The gigification of work". In: *Institute for the Future of Work* (2021). Available at: URL: `https://tinyurl.com/mryzrscv`.

[21] Binns, Rueben, Abigail Gilbert, Anna Thomas, Josh Simons (2020) 'Mind the Gap: The Report of the Equality Task Force' Institute for the Future of Work. Available at: `https://www.ifow.org/publications/mind-the-gap-the-final-report-of-the-equality-task-force`

highlighted the need for transparency in how AI systems are being used and how organizations go about AI risk assessments.[22]

Katie Shay (Cisco Systems Associate General Counsel & Director of Human Rights) also stressed the importance of transparency and accountability in AI, highlighting Cisco's Responsible AI framework[23]. This framework has five pieces that work together to operationalize Cisco's human rights approach to AI.

First, the company established a Responsible AI committee to drive the adoption of the principles of transparency, fairness, accountability, privacy, security, and reliability. The committee, comprised of executives across Cisco, reviews how the Responsible AI framework is being implemented.

Next, Cisco created a set of controls that engineers are required to use when developing Cisco products. The controls compel the engineers to complete a Responsible AI Assessment. This involves answering questions that help assessors determine the possible implications of a model, such as whether it could generate an output that results in a "consequential" decision affecting a certain group. If they find that a model might have a legal or human rights impact, the engineers are required to complete a more robust impact assessment.

Third, Cisco developed an incident management framework, where instances of potential bias in the AI are identified, reported to the governance committees, and remediated.

Fourth, Cisco engages with others in the industry to advance responsible AI.

Fifth, Cisco works with governments to monitor, track and influence emerging regulations and partners with research institutions working at the intersection of ethics and AI.

Dennis Hirsch (Professor and Director of the Program on Data and Governance at Moritz College of Law),[24] and Nicolas Vermeys (Associate Director of the Cyberjustice Laboratory and Professor at the Université de Montréal)[25] discussed managing AI risks from the perspective of law and policy. Professor Hirsch grounded his discussion of AI governance in the United States on an analogy between environmental harm and data harm. He noted that early environmental regulations took a first-generation approach that stipulated prescriptive rules with specific design standards — essentially a "command and control" model of regulation. Later, he noted a shift to a "second generation" regulation model that was more focused on setting targets that companies must meet and allowing companies to define their preferred approaches to meet those targets. Professor Hirsch claimed that the second-generation approach is what we are seeing in proposed AI laws in the United States, much of which is focused on performance-based regulation. These proposed laws emphasize the creation of systems that are fair and unbiased and allow organizations to figure out how to achieve those goals. This second-generation approach is planning-based (emphasizing the use of tools like Algorithmic Impact Assessments and audits to foster better outcome) and information-based (requiring algorithms to be transparent and explainable).[26],[27]

Professor Vermeys mentioned that one of the first forms of AI government regulation came out of the Canadian government in 2018. The Directive on Automated Decision, which only applies

[22]Gilbert, Abigail and Anna Thomas (2023) 'Good Work Algorithmic Impact Assessment: An Approach for Worker Involvement' Institute for the Future of Work Available at: https://www.ifow.org/publications/good-work-algorithmic-impact-assessment-an-approach-for-worker-involvement

[23]Cisco. "The Cisco Responsible AI Framework". In: (2022). URL: https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/cisco-responsible-artificial-intelligence-framework.pdf.

[24]Moritz College of Law: https://moritzlaw.osu.edu/

[25]Cyberjustice Laboratory: https://www.cyberjustice.ca/en/

[26]Dennis D Hirsch. "That's unfair-or is it: Big data, discrimination and the FTC's unfairness authority". In: *Ky. LJ* 103 (2014), p. 345. URL: https://uknowledge.uky.edu/cgi/viewcontent.cgi?article=1073&context=klj.

[27]"That's Unfair! Or Is It? Big Data, Discrimination and the FTC's Unfairness Authority". In: *Kentucky Law Journal* 103 (2015). Ed. by Dennis Hirsch.

to federal public bodies, requires those using AI to perform Algorithmic Impact Assessments, ensure transparency by providing explanations of AI systems and how they work, and perform quality assurance checks by conducting risks assessments during the development cycle of the system.[28] Internationally, legislators are beginning to develop regulation that focuses on high-risk AI systems—those systems used in areas that are so critical or dangerous that AI should either not be used or should be used with extreme caution. For example, Amazon's use of an AI hiring tool would be considered high risk, given the harm caused by the system's learned gender bias. The Artificial Intelligence Act in the EU,[29] for example, also classifies applications of artificial intelligence by risk and regulates them according to the risk level. This type of regulatory scheme is precautionary in that it requires those responsible for high-risk systems to take preemptive measures to identify and mitigate risks that could result from the use of the system.

While there is proposed AI legislation in the United States, Professor Hirsch noted that the law is not currently developed to the point of providing organizations with a compliance incentive for tackling problematic AI. Nevertheless, businesses are already taking active steps to identify and address the risks and threats created by their own use of AI. Businesses know that addressing these risks serves their commercial interests, such as reducing regulatory risk, building trust with customers, bolstering their reputations, and achieving a competitive benefit. Given that the management side of responsible AI governance is predominantly using this sustainability model, Professor Hirsch recommends we should attempt to design regulations to support this growing sustainability mindset in responsible AI management.

Finally, speakers discussed whether existing anti-discrimination and tort liability laws would provide sufficient measures to protect against AI risks. Panelists believed that, while this existing body of law does apply to AI/ML, they were not written with AI/ML in mind, so there are gaps.

Professor Hirsch suggested that since one of the main questions we need to answer with respect to AI is what is fair and what is unfair, the Federal Trade Commission's (FTC) "unfairness authority" may be useful. The unfairness authority allows the FTC to make case-by-case determinations on the fairness of certain practices, which ultimately builds into something like common law. This may be valuable in the rapidly-changing AI space, where it is difficult to develop prospective rules.

3.3 Panel 3: How well do current regulations/policies address these challenges?

For the third panel, Peter Chapman (Associate Director and Tech and Human Rights Lead, Article One Advisors)[30] began his discussion of AI regulation through the lens of the United Nations Guiding Principles (UNGP)[31] which emphasize the need for businesses to show that they respect human rights. He pointed out that technology will have both positive and negative impacts on people, and that businesses need to act with due diligence to intentionally advance human rights. He remarked that a commitment to governance of human rights risks can help inform responses to challenging questions, such as: What remedies should companies provide when their program does harm? What sorts of oversight or analysis must their algorithms go through?

While it is relatively simple for a company to release ethical AI principles, which are the foundation for norms, actual risk management, governance and oversight is much more challenging. Mr. Chapman believed that AI risk management can be informed by companies from one sector

[28] Canadian Directive on Automated Decision-making: `https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592`

[29] EU Artificial Intelligence Act: `https://artificialintelligenceact.eu/the-act/`

[30] Article One Advisors`https://articleoneadvisors.com/`

[31] UN. "Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework | OHCHR". in: (2012). URL: `https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights`.

considering the methods used in other sectors. This cross-sector approach may lead to more robust oversight and to more uniform norms for legislators to consider as a basis for regulation. Regulators should look to the UNGP and various existing company's work to build more effective AI governance. Mr. Chapman concluded by emphasizing that companies should map out the rights of users more clearly so that they can better clarify and detect when harm has been done.[32]

Brenda Leong (Partner, BNH.AI[33]) explained how algorithm auditing — similar to those in the financial market — could be a valuable tool for businesses to ensure their algorithms are ethical.[34] She noted that the current quality of AI audits is similar to those for the financial market in the 1950s, which lacked wide adaptation, an audit enforcement body, and accuracy and validity. Ms. Leong emphasized that audits are a tool, not the entire solution. Audits can be over-trusted and lead to normalizing bad programs when there is no second level of enforcement. They are a necessary but insufficient part of the framework that can check for risks, proper datasets, and development holes. She stressed the need to implement systems to ensure products are performing as they are intended when interacting with people, and appropriately delegating responsibility when there are problems. Audits should be an essential component of an overall best practices for AI use.[35]

Reva Schwartz (Research Scientist/Principal Investigator for AI Bias at the National Institute of Standards and Technology (NIST)) discussed the NIST AI risk management framework.[36] She noted past assessments were brittle because they approached a socio-technical problem only from a technical perspective. Now, NIST is working on incorporating a holistic test that can look more at impacts rather than errors. It is also focusing on involving humans in the assessment process in order to address AI's socio-technical aspects. Ms. Schwartz recommends using trustworthiness and intended audience as standards for AI systems. For trustworthiness, organizations should look at reliability and validity at a minimum and then push for safety, security, explainability, privacy, fairness, accountability, and transparency. She also highlighted NIST's recently published playbook on "how to make your programs trustworthy." Ms. Schwartz suggested that the AI community should adapt this framework to their specific industry requirements.[37]

Jessica Smith (Technology Policy Manager, Ofcom) described both the potential benefits and risks of algorithms being used in the communications sector, from filtering scam calls to recommender systems serving online users illegal or harmful content. Ofcom[38] is the UK's communications regulation, and regulates regimes from post to telecoms, and in the future online safety. The UK doesn't have a single AI regulator; however, AI risks can be addressed through existing regulations, ranging from data protection to competition. Ofcom is currently building internal technical capabilities to effectively govern AI algorithms through a Data Group which will provide advice for policymaking and model systems.

The UK and EU have both proposed regulations for AI. The UK has outlined a non-statutory framework which is based on the OECD's AI principles and includes fairness, safety and trans-

[32]United Nations. "Towards an Ethics of Artificial Intelligence | United Nations". In: (2018). URL: https://tinyurl.com/2vv7ysbc.

[33]BNH.AI: https://www.bnh.ai/

[34]Haziqa Sajid. "How to perform an AI Audit in 2023 - Unite.AI". in: (2023). URL: https://www.unite.ai/how-to-perform-an-ai-audit-in-2023/.

[35]BNH.AI Firm: https://www.bnh.ai/our-work

[36]NIST. "NIST - AI Risk Management Framework". In: (2023). URL: https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF.

[37]NIST. "NIST AI RISK Management Framework - Playbook". In: (2023). URL: https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.

[38]Ofcom homepage: https://www.ofcom.org.uk/home

parency[39]. UK regulators will seek to consider these principles within their regulatory remits. The European Commission (EC) has drafted a statutory regime that creates 'tiers' of risk and will place obligations on AI firms based on the level of risk that their product or service poses. For example, the use of facial recognition in public spaces will be banned, while AI-driven recommender systems that could influence voters have been classed as 'high risk' and will require firms to take measures in order to demonstrate appropriate safety.

Beyond regulation, Jessica mentioned several additional tools to help govern AI. First, she discussed the UK's Digital Regulation Cooperation Forum[40], where digital regulators collaborate to undertake research and share best practice on a range of issues. Second, she pointed to UK regulators who have been building their internal technical capabilities in order to respond to new AI developments, including for example, to undertake algorithmic assessments. Finally, she noted the UK's AI Standards Hub[41] which explores how regulators and firms can use relevant standards to support compliance with regulation.

3.4 Workshop Findings

The final session of the Problematic AI conference was a dynamic, interactive workshop that built upon on insights from the previous panel sessions. This workshop included panelists and in-person conference participants. It's goal was to propose the best way forward, given the findings from the previous panel discussions.

One of the first points raised during the workshop was to consider changing the conference title to "Problematic Humans," instead of "Problematic AI." The technology is not, by its nature, problematic, but misusing it—including operating outside of its intended context—can result in erratic, problematic behavior. As an example, Microsoft Word, a program written in a classical computer programming approach, does not alter its performance based on whether it is used in drafting legal papers or preparing medical notes, even though the content may be quite different. However, an AI model trained on legal texts may have significant challenges preparing medical notes. Some of the root causes of "Problematic AI" stem from improper training, or from applying a model to a context for which it was not trained.

In a related discussion, workshop attendees highlighted discrepancies on how AI systems are assessed. First, there is the question of comparative measures. Do we determine that an AI is satisfactory if it performs at least as well as a human? If not, then what measure do we use? Second, how do we measure model performance? Do we measure model performance solely as an autonomous entity or as part of a system of other technical and human components? Generally, panelists preferred considering the model as a component of the broader system, but such an approach makes testing and assessment much more complex. To audit models, we may need context-specific tools that use measurements tuned for the models' environment and purpose.

The workshop's discussion than moved to regulation. Participants agreed that regulation also must consider a model's context. For example, there are general rules of economics and fairness for corporate behavior, but there are also very specific regulations for industries in socially sensitive markets, such as Healthcare and Finance. Some attendees believed that AI regulation should follow a similar approach that can apply extra caution and oversight based on an industry's importance to society and the government.

[39] UK. "UK AI Rulebook". In: (2022). URL: https://tinyurl.com/bddac93y.

[40] UK. "The Digital Regulation Cooperation Forum - GOV.UK". in: (2023). URL: https://www.gov.uk/government/collections/the-digital-regulation-cooperation-forum.

[41] UK. "AI Hub: The One Place for Everything AI (Cloud Next '19 UK) - YouTube". In: (2020). URL: https://www.youtube.com/watch?v=CTjVtlxClck.

Another regulatory issue was that, without regulation, many technology companies are establishing their own principles and policies to govern AI's use in products and services. While commendable, this approach may have drawbacks. First, without clear external performance goals, private regulation could lead to great inconsistency as each company adopts its own perspective of what is "good" and "right." Second, as a promising cutting-edge technology, AI is widely seen as performing better on many tasks than human-managed systems. But, without external regulation, some organizations may fail to set important limits on AI's use (e.g., judging capital cases).

The participants agreed that the discussion of these topics should be continued and endorsed the idea of a 2024 symposium, which will also be supported by this year's conference sponsors, COVA CCI and the University of Montreal's CyberJustice Laboratory. Stay tuned for the date of the 2024 continuation of "Problematic AI: Finding the Best Way Forward."

4 Conclusion

Model misuse is one of the main factors that can make AI problematic and can occur in multiple scenarios. For example, model misuse can occur when AI technology is applied to an unsuitable, problematic area, such as allowing the system to have unacceptable control over human life, health, prosperity, and happiness. It may also occur when applying an AI model to a domain for which it was never designed (e.g., using ChatGPT to prepare legal briefs[42]). Finally, model misuse could occur when AI developers publish models without the proper oversight and controls, thereby creating models prone to producing erroneous results. Without proper system design and development controls, AI systems will likely perform poorly in production even though they perform perfectly during testing.[43,44] Panelists and attendees agreed that regulation would continue to be an important step to improve AI's trustworthiness, especially in the eyes of the public.

One approach to government regulation is to limit AI's use within problematic areas and then relax those limits once AI proves to be reliable and safe. In the US, Illinois (BIPA),[45] and Virginia (HB2031),[46] and Washington State[47] have implemented limitations on facial recognition. Maryland[48] requires consent for using facial recognition during job interviews. Oregon restricts law enforcement from using facial recognition on body cameras[49]. More regulation is pending to address broader AI applicability and appropriate use. Draft regulations are defining a set of 'high risk' domains that then require increased scrutiny and governance.

But how can AI risks be identified more readily and accurately? Panelists suggested that AI development teams should engage key stakeholders to identify potential AI risks early in system development. Developers should then work to mitigate those risks during implementation and then

[42]SD New York US District Courts. "MATA v. AVIANCA INC (2023) | FindLaw". In: (2023). URL: https://caselaw.findlaw.com/court/us-dis-crt-sd-new-yor/2205760.html.

[43]Marty J Wolf, K Miller, and Frances S Grodzinsky. "Why we should have seen that coming: comments on Microsoft's tay" experiment," and wider implications". In: *Acm Sigcas Computers and Society* 47.3 (2017), pp. 54–64.

[44]Carissa Véliz. "World view". In: *Nature* 615 (2023), p. 375.

[45]Illinois. "Illinois General Assembly - Full Text of Public Act 095-0994". In: (2008). URL: https://www.ilga.gov/legislation/publicacts/fulltext.asp?Name=095-0994.

[46]Virginia. "Bill Tracking - 2021 session > Legislation". In: (2021). URL: https://lis.virginia.gov/cgi-bin/legp604.exe?212+ful+CHAP0537.

[47]Washington State Legislature. "Chapter 43.386 RCW: FACIAL RECOGNITION". in: (2021). URL: https://app.leg.wa.gov/RCW/default.aspx?cite=43.386&full=true.

[48]Maryland. "Labor and Employment – Use of Facial Recognition Services – Prohibition". In: (2020). URL: https://mgaleg.maryland.gov/2020RS/chapters_noln/Ch_446_hb1202T.pdf.

[49]Oregon. "ORS 133.741 - Law enforcement agency policies and procedures regarding video and audio recordings". In: (2023). URL: https://oregon.public.law/statutes/ors_133.741.

monitor the model's performance within the whole system after it is released. This approach first gives developers new insights about potential risks and then ensures full life cycle monitoring to validate that the whole system continues to mitigate risks. If Amazon had taken this approach when employing AI in screening employment candidates, it might have classified its use as high risk and required the development team to perform more rigorous impact assessments and stakeholder engagement[50]. However, an important point about regulation highlighted by the Amazon example is that AI models are already governed by existing regulations. In Amazon's case, its AI model was already governed by Title VII of the Civil Rights Act of 1964[51] SEC. 2000e-2. [Section 703] and the General Data Protection Regulation for consent.[52] The panelists emphasized that, while there may be some gaps in regulatory regimes, AI technologies are already governed by explicit and well-established regulations. Where there are gaps, one panelist suggested that the FTC's "unfairness authority" may be a good regulator because it can quickly adjust and apply relevant rules in the rapidly evolving AI space.

Panelists suggested other tools that can help companies to determine and manage AI risk such as third party Algorithmic Impact Assessments (AIA) and AI Audits. In fact, the panelists noted that industry is already managing AI risks. For example, Microsoft and Cisco have demonstrated their commitment to managing AI risks through their documented principles and controls. Cisco described its Responsible AI framework that included the principles of transparency, fairness, accountability, privacy, security, and reliability, along with the internal corporate education, controls, and incident management. As with most corporate responsible AI work, this was done before existing regulation and was mainly inspired by industry activity and customer expectations.

A new computing paradigm is maturing that has great promise but also carries possible peril. More than ever, we need to be explicit about personal data use, model assessment, establishing general and industry-specific standards, and providing the tools and regulations that measure and transparently communicate model performance against those expectations. However, there remain significant challenges. One is measurement. Do we measure model performance against the human it is replacing/supporting? Do we measure model performance based on historical trends, or do we strive for different outcomes that may achieve desired equity or diversity goals? Another challenge is proving AI trustworthiness. An important factor influencing AI's trustworthiness is public perception. One demonstration of the public perception challenge is the fact that 1.35 million people die globally from car accidents each year with little change in public perception of automotive safety[53]. However, when the media covers one self-driving car accident, public perception of AI trustworthiness drops significantly.[54] For self-driving cars, at least, public perception seems to be less based on actual data but more on media coverage and the distrust of AI for certain tasks.

Overcoming these challenges requires a cross-discipline approach, much like the one we employed for this conference. Representatives from academia, law, industry, and policy-makers need to collaborate in order to realize the benefits of this remarkable technology while minimizing harm.

[50] Akhil Alfons Kodiyan. "An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool". In: *Researchgate Preprint* (2019), pp. 1–19.

[51] US-Congress. "Title VII of the Civil Rights Act of 1964 - U.S. Equal Employment Opportunity Commission". In: (1964). URL: https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964.

[52] EU. "What are the GDPR consent requirements? - GDPR.eu". In: (2018). URL: https://gdpr.eu/gdpr-consent-requirements/?cn-reloaded=1.

[53] Who In December. "Global status report on road safety 2018". In: (2018). URL: https://www.who.int/publications/i/item/9789241565684.

[54] Kareem Othman. "Public acceptance and perception of autonomous vehicles: a comprehensive review". In: *AI and Ethics* 1.3 (2021), pp. 355–387. ISSN: 2730-5961. DOI: 10.1007/s43681-021-00041-8. URL: https://doi.org/10.1007/s43681-021-00041-8.