Ethics and Governance of Artificial Intelligence Initiative
Berkman Klein Center for Internet & Society at Harvard University
MIT Media Lab at the Massachusetts Institute of Technology

Algorithms and Justice

Government institutions around the globe are beginning to explore decision automation in a variety of contexts, from determining eligibility for services; to evaluating where to deploy health inspectors and law enforcement personnel; to defining boundaries around voting districts. Use cases for technologies that incorporate AI or machine learning will expand as governments and companies amass larger quantities of data and analytical tools become more powerful.

The criminal justice system offers valuable insight into government use of algorithmic technology. With fallible judges, juries, and lawyers, that system has been rightly criticized for inconsistency and for perpetuating practices that disproportionately harm marginalized groups. Support for reexamination of detention practices has grown in recent years, as reformers and state institutions alike seek to control costs, manage overcrowded prison systems, and address disparate impacts of incarceration.

To the extent they inject clarity and precision into bail, parole, and sentencing decisions, algorithmic technologies may minimize harms that are the products of human judgment. Conversely, the use of technology to determine whose liberty is deprived and on what terms raises significant concerns about transparency and interpretability. We must consider both legal and ethical issues and engage in rigorous testing and evaluation to ensure adoption of algorithmic tools satisfies notions of procedural and substantive fairness and does not reinforce institutional biases.

As its name suggests, the criminal justice system is not a unified construct but a series of interconnected processes, with multiple entry points and stages of evaluation. Technology may help to balance among goals of deterrence, incapacitation, rehabilitation, restitution, and retribution, shedding light on what causes criminal behavior and appropriate responses thereto. Each stage in the development, procurement, deployment, and assessment of each technological tool raises distinct and essential questions that demand a multiplicity of approaches.

> **The Algorithms and Justice track explores ways in which government institutions incorporate artificial intelligence, algorithms, and machine learning technologies into their decisionmaking.** Our aim is to help the public and private entities that create such tools, state actors that procure and deploy them, and citizens they impact understand how those tools work. We seek to ensure that algorithmic applications are developed and used with an eye toward improving fairness and efficacy without sacrificing values of accountability and transparency. Our work begins with a focus on the United States, while developing more generalizable lessons and best practices.

In undertaking this work, the **Berkman Klein Center** will draw on its rich history of delving into hard law and policy questions via research and engagement with government actors and innovators. The **MIT Media Lab** will draw on its rigorous application of research methods, technical expertise, and outside-the-box thinking.

Challenges that we seek to address in our work include:

- **Transparency**: The law traditionally places great importance on transparency in the workings of government and—in particular—in the administration of the justice system. Development processes and methodologies can be opaque, and jurisdictions do not always provide access to data that allows for oversight of technology-enabled decisions. For this reason, we are building a database of the most common risk assessment tools used in the United States, to illuminate the methodologies and limitations of such tools.

- **Bias**: In the United States and elsewhere, historically-marginalized groups are often over-represented in incarcerated populations. Algorithmic systems trained on historical data must therefore confront inherent biases. Existing assessment tools approach the legacy of unequal outcomes in different ways. But little is known about the effectiveness of their methods, and there exists considerable debate about the extent to which these factors (and their proxies) can be isolated.

- **Due process**: Use of technology in the criminal justice system has the potential to upend centuries-old conceptions of due process and force debates about adapting norms to suit the digital age. The outcome of those debates will hinge on whether new challenges are analogous to past ones (from which we can learn), or whether they represent an existential crisis for the judicial system (demanding a reimagining of criminal liability and punishment). Addressing these issues requires an interdisciplinary approach, translating concepts of justice and fairness between lawyers and policymakers (on the one hand) and technologists (on the other). Importantly, by partnering with local communities, we are working to demonstrate that the judicious and timely application of technology in can actually improve social service deliver, reduce interactions with the judicial system, and better advance the core motivations of the justice system as a whole.

- **Competing priorities**: Undeniable tensions exist among commercial interests of those that build and sell technology; cost-management interests of government officers that procure tech tools; and societal interests of citizens seeking to preserve norms and values around fairness, justice, and accountability. This is why it is critical that we are working closely with state attorneys general, and court administrators to develop resources to help manages these competing priorities.

- **Interpretability and Dynamic Adaptation**: Most government decisions (including decisions about incarceration) are accompanied by legal analyses and justifications. But algorithms may not offer explanations that laypersons can understand, and the dynamic nature of machine learning might yield different results in one case today and a similar case tomorrow. It is vital that we ensure results are interpretable and that any inconsistencies can be explained. To advance this, we are supporting an interdisciplinary group of lawyers and computer scientists who are working on a framework to help policymakers understand the advantages and limitations of algorithmic explanations, so that policies can be grounded in technically feasible approaches.

Pillars of Impact

In building solutions that address these challenges, our institutions are making a series of investments, most significantly in:

1. Launching a **database of risk assessment tools**, including information about where and how they are used; development and validation practices; and legal, legislative, and/or media responses to their use.
2. Working with the town of Chelsea, MA to improve and expand programs designed to **identify and remediate the root causes** of crime through data-sharing and advanced algorithmic techniques, rather than generating "predictions" based on correlated factors, and creating best practices for similar programs nationwide.
3. **Work with jurisdictions** to develop best practices around the procurement and deployment of risk assessment tools to govern their adoption and create frameworks for assessing impact.

For more information, please contact Chris Bavitz at cbavitz@cyber.harvard.edu or visit https://cyber.harvard.edu/research/ai.