Comment on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

Andrew P. Wheeler, PhD andrew.wheeler@crimede-coder.com

2024-05-16

Summary:

- I provide a working definition of AI, focusing on how the tool will be used repeatedly in practice
- A key component of trustworthiness of a model is the data used to construct it.
- Safety is defined by how technology will be used in practice; the same technology can be used in safe or unsafe ways
- NIJ should expand its funding portfolio to promote original research into the development of AI models.

Table of Contents

Defining Artificial Intelligence

Artificial intelligence (AI) in the context of technological development needs a reasonable working definition. I propose the following:

> Systems that use historical data to generate automated predictions, that are used repeatedly to make or aid actions a criminal justice agency takes.

Here I focus on aspects *of how the technology will be used*, and less so on drawing a bright line as to what does (or does not) constitute AI.

Many current applications of AI are intended to be replacements to particular human decision making, with examples such as automated facial recognition in place of manual review, predictive policing in place of ad-hoc hotspot maps created by a crime analyst, automated report writing in-place of officer written narratives, etc.

A common component of these examples is the repeated aspect. Being able to replace repetitive tasks both have the most upside for criminal justice agencies, as well as the greatest potential for harm in their application. Articulating that the predictive systems use historical data will be elaborated on to its importance in the subsequent section.

Components of Trustworthy Development of AI

I placed in my definition *historical data* – as a developer of predictive models, I believe understanding the nature of the data used to create predictive models is the most important component for outsiders to evaluate the relevance of an AI model. The nature of models (such as the whether the predictive model is tree based or uses a deep learning architecture), can in my opinion be reasonably kept secret protecting the intellectual property of developers.

Limitations and flaws in the original data however can fundamentally change how one uses the particular model based predictions in practice.

For one example, imagine a company builds a facial recognition algorithm, in which it is trained on pairs of 50% matches and 50% not-matches of data. In reality, when applying this predictive model in the population of data (in which one takes one image and scans many thousands of other images), the rate of matches will be far less than 50%. Thus even if the model outputs a 99% match, in reality one should down adjust that predicted probability. If the company however never released details on the training of the original algorithm, one would not be able to know the context of what a 99% match probability means in practice.

For a second example, imagine a company suggests to help write incident narratives to save officer time. Generative text models, such as ChatGPT, are fundamentally models that use historical text to predict future text sequences. Their outputs are limited to what they have seen in historical examples. If a company

used historical narratives, but those narratives were biased to discount victim reports (see the work of Rachel Lovell in rape incident reports for example) the automated tool may help reinforce such historical biased narratives.

I do not take these examples as exhaustive – but I believe reasonably disclosing the nature of data used to create an AI model will be a necessary component for outsiders to evaluate its trustworthiness.

Components of Safe Use of AI

Safety is dictated *by how the technology will be used in practice.* Consider a spatial predictive policing system; one police department may use the technology to conduct intensive policing into the areas it prioritizes, whereas another uses the tool to help formulate community oriented policing initiatives to reduce crime. The former has a much greater risk of infringing civil liberties and having negative externalities in terms of harm to the community than the latter, even though they are based on the same underlying technology.

Safe use of AI therefore needs *for any particular technology to specifically define how it will be used in practice*. Scope creep in the application of a tool, such as expanding the use of automated license plate readers from its original role of flagging stolen cars to a more general tool that can monitor the whereabouts of individuals, can result in unintended harms.

The Role of NIJ in Promoting AI

I believe the best way forward for NIJ is to fund original research into creating open source AI models, as well as specifically funding outcome research on applying those technological tools in practice.

NIJ's current portfolio of funding research in collaboration of criminal justice agencies is strong, and requires no dramatic change in priorities. It may be reasonable to have specific applications in terms of applying AI tools in criminal justice practice, but I believe this could be quite easily expanded to being a component of other topical applications.

Where I believe NIJ's portfolio of research is lacking is funding of original development of such models to begin with. NIJ is in a prime position to encourage researchers to develop open source models to tackle some of these challenges across criminal justice, and these include applications such as predictive early warning systems for police officers, generative AI applications to help with report writing, automated evaluations of video and audio inputs, to name a few.

By funding such original work, NIJ accomplishes two things. One, it establishes open source implementations of models, and Two specific reference datasets on which those models were developed. This provides solid ground for any future

(public or private) implementation to be evaluated. I believe this will ultimately improve the quality of such AI models in practice, without substantively hindering private development in the long term. Without such standard benchmarks and open source implementations however, there is very little to gauge the trustworthiness of any closed source implementation.

Generating such models takes significant time and effort. Only funding research that has a specific implementation component substantially hinders the ability of researchers to undertake such work. That does not mean models need not ever be evaluated in how they are used in practice, just that development of AI models and tools is worthy to fund in and of itself.

About CRIME De-Coder

CRIME De-Coder LLC was formed by Andrew P. Wheeler, PhD in 2023. **CRIME** De-Coder was formed as a way to help police departments tackle data science related problems that many agencies need outside help with.

The name **CRIME** De-Coder originates from how Dr. Wheeler de-codes the complexity of computer programming and statistical analysis to help police agencies with their problems.

Dr. Wheeler has a PhD in criminal justice from SUNY Albany (2015), and has worked with police departments and different criminal justice agencies across the United States. During his graduate studies, he worked as a crime analyst at Troy, NY police department. Afterward, he was a professor of criminology at the University of Texas at Dallas, and was finally employed as a data scientist in the private sector. He currently has over 35 peer reviewed publications that have been cited over 1000 times. These include applications of machine learning and fairness in the application of criminal justice operations.

You can view the portfolio of **CRIME** De-Coder's work at crimede-coder.com