William & Mary Law School
William & Mary Law School Scholarship Repository

Popular Media | Faculty and Deans

5-17-2022

Problematic AI — When Should We Use It?

Fredric Lederer

Follow this and additional works at: https://scholarship.law.wm.edu/popular_media

 Part of the Artificial Intelligence and Robotics Commons, Criminal Procedure Commons, and the Military, War, and Peace Commons

Copyright c 2022 by the authors. This article is brought to you by the William & Mary Law School Scholarship Repository.
https://scholarship.law.wm.edu/popular_media

Problematic AI — When Should We Use It?

May 17 Written By Fredric Lederer

One of the most interesting and important developments in the ongoing evolution of technological tools, Artificial Intelligence (AI) is a largely misunderstood, advertising-hyped technology that is already changing our lives. Public reaction to AI use varies widely. To many, it's simply the same as saying, "produced or affected by computers." To some, it's a near magical technological innovation leading, for example, to autonomous, "self-driving" vehicles. And to others, it suggests deadly killing machines — "Killer-bots" that could supplant and destroy humanity.

Strictly speaking, according to Dr. Karl Branting of Mitre, "Artificial Intelligence" or AI can be described as "the field that seeks technology capable of human-like reasoning, perception, and control." Machine learning is a form of AI that modifies its own programming as it "learns" via exposure to new data. Accordingly, unless otherwise differentiated, my use of "Artificial Intelligence" or "AI" refers to forms of machine learning. For machine learning to function correctly, the AI program software — the "algorithm" — must be written correctly. Then, in most cases, it must be "trained." Training means that the algorithm is exposed to large amounts of accurately labeled data. AI permits computerized analysis of these vast

amounts of data in order to discern potential causation — what facts or circumstances cause a given result. With that determination, AI systems can identify patterns, make predictions, and, if so designed, implement those predictions.

AI can vastly improve life. Consider the January 2022 traffic stoppage on Interstate 95, which trapped drivers on the highway, in some cases for up to 24 hours, because of snow. One can imagine a future AI system analyzing probable real world weather conditions, correlating them with available snow removal resources and "known" data about how drivers react to past similar weather conditions, and then advising or implementing a highway closure.

However, the nature of AI can also be problematic. No matter how it is expressed, the result of an AI data analysis is probabilistic. No matter how certain an AI result seems, the answer is always a probability, one that is dependent on data. Further, because AI modifies its own programming as it analyzes new data, we do not always know why an AI reaches a given result. When algorithms are trained with or use incorrect or biased data, the algorithm produces biased results. For example, thanks to bias and thus error, a number of cities have banned facial recognition use for law enforcement. Technology companies using AI to select good job candidates have discovered that using the resumes of current employees, mostly male, taught the AI to select men.

Even though training an AI with inaccurate data is a major concern, If training were the only problem, we likely could refine our process and produce accurate AI systems. However, the key to machine learning is that the algorithm "learns" as it operates; it alters its own programming to correspond with the data it uses in operation. If the algorithm is designed to encourage people to eat healthy foods, for example, and people report to the algorithm (or post on social media that is "read" by the algorithm) that they are healthier by eating fast food, the algorithm will recommend fast food unless its designers have written its programming in a manner that prevents such a possibility. Access to the Internet permits the use of large amounts of data, but often that data is inaccurate — at best. In 2016, an experimental Microsoft "Bot" designed to mimic a teenage girl received access to Twitter. "Less than a day after she joined Twitter, Microsoft's AI bot, Tay.ai, was taken down for becoming a sexist, racist monster." Gaining more than 50,000 followers, Tay.ai mimicked "her" followers. Therefore, the use of technology to assist in decision-making or to make important decisions without major human involvement has often been troublesome.

The question, then, is how society ought to deal with AI, a promising technology that at times is certain to malfunction. There may be uses in which accuracy and avoiding harmful behavior are so important that we must not tolerate any potential AI error. If there is a serious risk that an AI driven vehicle will "choose" to run down humans, for example, we must not use AI for that purpose unless we can install adequate safeguards.

Usually, those who are troubled by current or future AI use have two primary concerns. First, in some areas AI decision-making should not be a substitute for fundamental decisions only humans should make. Second, for most AI uses, the problem is that in light of *how people behave,* AI has a strong likelihood of yielding erroneous results that will be assumed to be correct. Evolving ethical standards and algorithmic justice may be useful; however, given the nature of AI, it's not helpful to say, "Do no harm."

We live in an imperfect world, one that contains many types of bias. The key problem in using AI algorithms is that AI tends to reflect the real world's bias. Given that our goal is — or ought to be — making our world better, which is to say not just more efficient or less expensive to live in, but also fairer, and more equitable, I believe the following three-part framework to examine AI's usefulness can be of value.

> ***Will the proposed AI use most likely be no worse than current reasonable human analysis and decision-making?***
>
> ***Does the proposed AI use reflect reasonable efforts to eliminate inaccuracy and bias?***

> ***Is there a reasonable chance that if the AI use is allowed to evolve, it will become fairer and more accurate than human efforts?***

These recommendations embrace the classic cliché, "the perfect is the enemy of the good." In dealing with human decision-making, comparisons ought to be made using reality — the average, normal person — rather than a near-mythical, all-knowing wise person who doesn't exist.

Two AI uses that well illustrate my analytical framework are the use of AI to predict recidivism for pretrial release and sentencing in criminal cases and the use of AI-controlled weaponry.

Predicting Future Criminal Misconduct

Sometimes, we need to determine the likelihood that a person will commit a crime. When persons are arrested, they go before a judge or magistrate who must determine the conditions, if any, by which they can be released while awaiting trial. In theory, absent a statute or court rule permitting preventive detention to avoid likely future misconduct, the question should be what condition or combination of conditions will ensure that the defendant will show up for trial. However, in addition to the numerous jurisdictions with formal preventive detention procedures, most judges take into consideration the possibility that the accused may commit another crime pending trial. This consideration can result in pretrial detention or sharply increase the amount of bail, quite possibly to such a level that the accused cannot secure release. How does a judge, who does not personally know the defendant, determine the likelihood that that person will commit a crime before trial, especially a crime of violence? The same question faces a sentencing judge who must determine what punishment to mete out after a defendant is convicted. If the judge is concerned that a defendant is likely to reoffend, the judge may wish to impose a more severe sentence to protect the public.

A number of organizations now offer "Risk Assessment" tools — non-machine learning technology that is intended to predict the likelihood of future criminal conduct. Originally intended to help with post-incarceration release decision-making, this technology is now used in both pretrial release decisions and sentencing. A judge or court official inputs basic data about the defendant, and the algorithm provides what amounts to a probability score prediction of future crime. The technology does not make the actual pretrial release or sentencing decision. The judge does that, but only after taking the AI prediction into consideration.

The use of this technology was challenged in a well-known 2017 Wisconsin State Supreme Court case involving sentencing, *State v. Loomis*. In *Loomis*, the defendant asserted, among other things, that his sentence violated due process because the court employed a commercial technology product, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), to predict his likely recidivism. As the Court explained:

> COMPAS is a risk-need assessment tool designed by Northpointe, Inc. to provide decisional support for the Department of Corrections when making placement decisions, managing offenders, and planning treatment. The COMPAS risk assessment is based upon information gathered from the defendant's criminal file and an interview with the defendant.
>
> A COMPAS report consists of a risk assessment designed to predict recidivism and a separate needs assessment for identifying program needs in areas such as employment, housing and substance abuse. The risk assessment portion of COMPAS generates risk scores displayed in the form of a bar chart, with three bars that represent pretrial recidivism risk, general recidivism risk, and violent recidivism risk. Each bar indicates a defendant's level of risk on a scale of one to ten.
>
> As the Presentence Investigation Report (PSI) explains, risk scores are intended to predict the general likelihood that those with a similar history of offending are either less likely or more likely to commit another crime following release from custody. However, the COMPAS risk assessment does not predict the specific likelihood that an individual offender will reoffend. Instead, it provides a prediction based on a comparison of information about the individual to a similar data group.

Loomis attacked the use of the COMPAS prediction on other grounds as well, including that he was not allowed to inspect the computer code itself because COMPAS' manufacturer, Northpointe, Inc., invoked the trade secret privilege to keep it secret. The Court opined that "Although Loomis cannot review and challenge how the COMPAS algorithm calculates risk, he can at least review and challenge the resulting risk scores set forth in the report attached to the PSI."

Ultimately, the Court sustained the use of the COMPAS risk assessment as a factor in determining a sentence, but made it clear that it was doing so because the judge was responsible for making the actual sentencing decision and the COMPAS assessment was only one factor the judge considered. The Court expressly noted ProPublica's analysis, which reported that COMPAS was racially biased:

> A recent analysis of COMPAS's recidivism scores based upon data from 10,000 criminal defendants in Broward County, Florida, concluded that black defendants "were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism." Likewise, white defendants were more likely than black defendants to be incorrectly flagged as low risk. Although Northpointe disputes this analysis, this study and others raise concerns regarding how a COMPAS assessment's risk factors correlate with race. [The interested reader may wish to consult Mission and Scope · Harvard Data Science Review, which concludes that COMPAS was not racially biased but very complicated in its structure.]

The Court opined that while any use of COMPAS would require the PSI to warn the user about concerns, providing the sentencing court with more information than would be available without the COMPAS assessment would still be helpful.

There are many criticisms of using risk assessment tools, not the least of which are that they are group-oriented and reflect past data. Moreover, those data fail to deal with the history and behavior of the specific person involved. In addition, as many Americans rely heavily on technology, they may well give undue weight to technological predictions, especially when they emanate from something called "Artificial Intelligence." This may be especially true as most lawyers and judges do not have technological backgrounds.

Why then did the Supreme Court of Wisconsin reject the *Loomis* challenges? The key can be found in its discussion of and preference for "evidence-based sentencing":

> It is helpful to consider Loomis's due process arguments in the broader context of the evolution of evidence-based sentencing. Wisconsin has been at the forefront of advancing evidence-based practices. In 2004, this court's Planning and Policy Advisory Committee (PPAC) created a subcommittee "to explore and assess the effectiveness of policies and programs ... designed to improve public safety and reduce incarceration."
>
> From that initial charge, Wisconsin's commitment to evidence-based practices and its leadership role have been well documented. Initially, a variety of risk and needs assessment tools were used by various jurisdictions within the state. In 2012, however, the Wisconsin Department of Corrections selected COMPAS as the statewide assessment tool for its correctional officers, providing assessment of risk probability for pretrial release misconduct and general recidivism.
>
> In [the prior Wisconsin Court of Appeals case] *Gallion* we warned of ad hoc decision making at sentencing: "Experience has taught us to be cautious when reaching high consequence conclusions about human nature that seem to be intuitively correct at the moment. Better instead is a conclusion that is based on more complete and accurate information..." We encouraged circuit courts to seek "more complete information upfront, at the time of sentencing. Judges would be assisted in knowing about a defendant's propensity for causing harm [and] the circumstances likely to precipitate the harm..."

> Concern about ad hoc decision making is justified. A myriad of determinations are made throughout the criminal justice system without consideration of tested facts of any kind. Questions such as whether to require treatment, if so what kind, and how long supervision should last often have been left to a judge's intuition or a correctional officer's standard practice.

Sentencing has long been recognized as a fundamental problem in criminal law and justice. Judges make important decisions with inadequate data. Judges are rarely provided with clear cut sentencing goals. When they formally exist, they usually are not ranked by priority of importance. In the absence of a statutory or rule requirement, it becomes inherently difficult, if not plainly impossible, to say what the "right" sentence is. Interestingly, in the allied area of pretrial release, the National Center for State Courts has reported:

> In an exhaustive analysis of a nationally representative sample of over 70,000 felony defendants on pretrial release from the seventy-five largest counties in the U.S. between 1990 and 2006, researchers found that historically judges have often released and detained the wrong people. Half of those detained had less than a 20% chance of re-arrest while an equivalent number of those released had a greater likelihood of committing a crime.

Inconsistent and arbitrary sentencing is commonplace. Disparity in sentencing by race is well recognized. It is not unusual for co-actors tried separately to receive highly inconsistent sentences in civilian courts. In a major effort to ameliorate the problem, Congress created the Federal Sentencing Guidelines, which attempted to provide some degree of consistent sentencing in the federal courts. In 2005, the Supreme Court held that they were not binding in *United States v. Booker*.

In most cases, judges need not explain their sentencing decisions. We have made progress in that legally and now culturally, we recognize that offenders should not have their sentences determined in whole or in part by race, ethnic background, religion, and the like — although we still have disparate sentencing in that regard.

In short, human decision-making in sentencing is inadequate. It is arbitrary, sometimes capricious. Often it defies explanation or adequate justification. Furthermore, this has been problematic for centuries. What of AI risk assessments? Of course, the COMPAS risk assessment used in *Loomis* was only one part of the sentencing data the judge used. Although human beings have changed little when it comes to sentencing, technology holds the promise of improvement. Its rapid pace of development is obvious. Today's "telephone" represents an example; smart phones have become data hubs, health monitors, communication centers, the source of personal videoconferencing, audio-video communications, and more. Had the Wisconsin Supreme Court held that the deficiencies in COMPAS violate due process, it likely would have signaled to commercial concerns that further development work would have been unwarranted financially. In light of its holding, along with those of other courts, companies can now continue working to improve their products notwithstanding the clear deficiencies in the current technology. There is a reasonable chance that eventually AI will prove superior to human judgment. When AI risk assessment, deeply flawed as it is, is compared with human behavior, it may not be so bad after all. More importantly, it probably will improve with time, while human ability appears to be largely fixed. This presupposes that the nature of risk assessments remains unchanged. It could be that augmented assessments could be used to help judges recognize their own biases and information shortcomings.

Applying the proposed AI uses framework to recidivism, we could conclude that:

> ***Will the proposed AI use most likely be no worse than current reasonable human analysis and decision-making?***
>
> Especially in light of what is understood about the limitations of AI risk assessments, courts that allow judges to use them but not completely defer to them are no worse than unaided human decision-making.
>
> ***Does the proposed AI use reflect reasonable efforts to eliminate inaccuracy and bias?***

> The limitations of technological risk assessments have been discussed extensively in both scholarly and public literature, and it would appear that work to improve the risk assessments is ongoing.
>
> ***Is there a reasonable chance that if the AI use is allowed to evolve, it will become fairer and more accurate than human efforts?***
>
> Given the nature of technological change, there is a reasonable probability that AI risk assessments will improve. The field is an evolving science.

This proposed AI use framework supports the Wisconsin Supreme Court's decision in *Loomis*, albeit in a more nuanced fashion.

Autonomous Weapons Systems

Modern warfare is lethal and fast. An attack can be launched with such surprise that notice, if any, is only a matter of minutes. Accordingly, there is great interest in potentially autonomous weapon systems that can remain ready for long periods, whether attended by people or not, and can react instantly when necessary. The fundamental moral objections to AI-controlled weapons are that machines ought not to "decide" to take a life, and that such weapons would inspire a new arms race. Rather, humans should either make the decision to fire or at least be "on the loop" and able to make the decision to withhold that permission from an autonomous weapon system. [*Note on terminology:* "on the loop" is different from "in the loop"; "in the loop" requires human decision-making while "on the loop" permits a human to stop the weapon.]

Setting aside the moral question, however, there often appears to be an assumption, stated or otherwise, that human judgment is possible on the modern battlefield and that our judgment is better than machine judgment. That assumption can be fallacious for two reasons: the nature of the modern high-tech battlefield in which decisions must be made faster than humans are capable of and the inherent fallibility of human decision makers, especially in combat.

Commendable efforts to constrain warfare have been with us for a long time and form the basis of the Geneva Accords, the Hague Conventions of 1899 and 1907, and the Geneva Protocol. Over the centuries, moral objections have been raised to various types of weapons. Gunpowder permitted killing at a great distance. Some claimed that was immoral because it removed the risk of harm from the target who otherwise would have to be personally faced. More recently, we have questioned the ethics of using nuclear weapons, "dumb" mines, and improvised explosive devices (IEDs).

Setting aside the moral issues surrounding war and weapons, objections to AI-controlled weaponry center on out-of-control weapons and the possibility that AI-controlled weapons systems could make decisions that harm or kill the wrong people. AI "Killer-bots" embody the fear personified in science fiction stories by Fred Saberhagen of AI "Berserker" starships that seek to destroy all life throughout the galaxy. Given that machines can malfunction, the need to be able to deactivate a weapons system seems obvious. The more difficult question is the degree to which, if at all, a human should have to authorize the operation of an AI weapon so that it cannot fire at a specific target without a real person's approval. Although it appears that the U.S. does not now have "lethal autonomous weapons systems," current U.S. policy does not require such human involvement, the requirement for a "human in the loop."

Modern warfare often is too fast for human judgment. Technology has enabled massive enemy attacks to be undetectable until the last minute. Defending against them is difficult. One response to this problem is the Phalanx CIWS Weapon System. As its manufacturer, Raytheon Missiles and Defense, defines it, the U.S. Navy Phalanx CIWS Weapon System "is a rapid-fire, computer-controlled, radar-guided gun that can defeat anti-ship missiles and other close-in threats on land and at sea." The system is:

> a fast-reaction, rapid-fire 20-millimeter gun system that provides U.S. Navy ships with a terminal defense against anti-ship missiles that have penetrated other fleet defenses. Designed to engage anti-ship cruise missiles and fixed-wing aircraft at short range, Phalanx automatically engages functions usually performed by separate, independent systems such as search, detection, threat evaluation,

> acquisition, track, firing, target destruction, kill assessment and cease fire... [The weapon fires] at either 3,000 or 4,500 rounds-per-minute...

Although a human being must turn on the system, it is otherwise automatic insofar as target engagement is concerned.

In addition to the need to choose targets in speed and data overload circumstances, technology often is used to compensate for inadequate troop strength and inadequate human decision-making. Consider "dumb" landmines. Placed on or under the ground to deny enemy access to large areas, mines detonate when a person steps on them, pulls a tripwire, or engages a similar device. Landmines not only kill innocent non-combatants during war but also often continue to be fatal risks for years after a war ends. Despite continuing protest, we use landmines largely because we don't have enough personnel to replace them. What about using AI systems to decide whether to fire a mine?

As of 2015, Samsung had manufactured the Samsung SGR-AI, which was installed in the Korean demilitarized zone. The system is a technological sentry that not only detects targets up to two miles away but also can function automatically using its own target detection system. Why might we want to use such a device in an automatic mode? Clausewitz is credited with the expression the "fog of war," the difficulty of knowing what is really happening in combat. If we were to use only human beings to decide whether to fire a weapon, who makes the fire/no fire decision? Unless we use the system now used for remote drone operations with drone pilots potentially thousands of miles away from the target, that person is likely to be an on-the-spot exhausted soldier or Marine with limited information trying to cope with numerous immediate requirements. How good will that person's decision be compared with a device that never gets tired, hungry, or distracted, and which has been programmed not to fire at what appears to be children or other non-combatants?

What of the remote drone pilots then? They are often stressed human beings with all that brings. They too make errors. The Defense Department has declared that the August 29, 2021 Kabul drone strike "killed 10 civilians, including an aid worker and seven children" despite the fact that the drone and its operator tracked the vehicle involved for hours.

If we compare AI-based weapons systems with human decision-making, it becomes apparent that AI systems may be able to function properly when human beings can't. Properly programmed with law of war constraints, AI systems may produce more lawful results than human military personnel can**.** That AI systems may err is clear. But we must compare that error to the probable human error. For an interesting discussion of this general topic see Mackenzie Patrick Eason, *Lethal Autonomous Weapon Systems: Reconciling the Myth of Killer Robots and the Reality of the Modern Battlefield*.

Returning to my proposed AI use framework and applying them to possible AI weapons systems:

> ***Will the proposed AI use likely be no worse than current reasonable human analysis and decision-making?***
>
> Human-controlled weapons system decisions are likely to be made by fallible people with inadequate information in circumstances in which their reasoning ability is adversely affected by combat conditions, inadequate information, and insufficient time — constraints less likely to apply to AI systems.
>
> ***Does the proposed AI use reflect reasonable efforts to eliminate inaccuracy and bias?***
>
> Yes. At least in the U.S., we make weapons systems that are intended to comply with law and practical, including political, needs.
>
> ***Is there a reasonable chance that if the AI use is allowed to evolve, it will become fairer and more accurate than human efforts are?***
>
> In light of the increasing analytical ability of AI systems and the ability to make faster-than-human decisions evaluating more data than a person can evaluate, the answer is "yes."

Conclusion

AI holds immense promise to improve our lives. Yet, we can predict that for at least the near and likely-midterm future, AI products will continue to be imperfect, sometimes in distressing ways. We can reject AI use, either by legal or political prohibition, or we can use artificial intelligence with suitable caution, safeguards, and understanding. Our policy should favor the latter option. It would be wrong to decide that human error and bias are pragmatically acceptable and reject technological help that might make society function in a fairer, more equitable, and generally better way. It is clear that AI may be so imperfect and so harmful to individuals or groups, especially when used by imperfect and possibly biased people, that we ought not to tolerate such harm. But, in deciding whether to tolerate a given amount of error and harm, we should ask whether AI holds sufficient promise for a better future at a cost equal to or less than current human decision-making. If so, we should tolerate that price.

About the Author:

Fredric I. Lederer is Chancellor Professor of Law and Director of the Center for Legal and Court Technology (CLCT) at William & Mary Law School. Professor Lederer received a B.S. from what is now NYU's Tandon School of Engineering, a J.D. from Columbia University School of Law, and an LL.M. from the University of Virginia. He was a Fulbright-Hayes Research Scholar at the Max Planck Institut fur auslandisches und internationales Strafrecht, in Freiburg, Germany. Professor Lederer's areas of specialization include legal technology, evidence, technology-augmented trial practice, electronic discovery and data seizures, the legal impact of the Artificial Intelligence ecosystem, criminal procedure, military law, and legal skills. He is the author or co-author of twelve books, numerous articles, and two law related education television series. He is a former prosecutor, defense attorney, trial judge, and law reform expert.

Technology Innovation in Social Impact SeriesSocial Impact Series