Committee For Justice
Holding Judges and Politicians Accountable to the Constitution

The Committee for Justice
1629 K St. NW Suite 300
Washington, D.C. 20006

(202) 270-7748
committeeforjustice.org
@CmteForJustice

Comments filed with Executive Office of the President, Office of Management and Budget:

Re: Guidance for Regulation of Artificial Intelligence Applications

DOCKET NO.: 2020-00261
SUBMITTED: MARCH 13, 2020

CURT LEVEY
PRESIDENT
THE COMMITTEE FOR JUSTICE

We at the Committee for Justice (CFJ) write to the Office of Management and Budget to comment on its draft memorandum entitled “Guidance for Regulation of Artificial Intelligence Applications.” Founded in 2002, the Committee for Justice (CFJ) is a nonprofit, nonpartisan legal and policy organization dedicated to promoting the constitutional principles of individual liberty, the rule of law, and limited government power. CFJ focuses, in part, on the preservation of these principles at the intersection of law and technology.

The aim of our comments is to assist the executive branch in creating policies that will continue and advance America's leadership in artificial intelligence technology and innovation while also promoting America's values. Specifically, our comments address two of the "Principles for the Stewardship of AI Applications" discussed in the draft memo: 1) fairness and non-discrimination, and 2) disclosure and transparency.

Our views on the draft memorandum are informed, in part, by the professional experience of CFJ's president, Curt Levey. After earning undergraduate and graduate degrees in computer science and before becoming an attorney, Mr. Levey worked as a scientist at HNC, Inc., an artificial intelligence startup company in San Diego, CA. There he built machine learning models known as neural networks for a variety of applications and traveled around the United States teaching people in various industries to build neural network models.

His experience as a scientist is particularly relevant to the draft memo's disclosure and transparency principle because, while at HNC, he invented and patented pioneering technology to provide explanations and confidence measures for the decisions made by neural networks. Moreover, his legal experience in civil rights—for example, managing the branch of attorneys and program analysts that develops Title IX policy at U.S. Department of Education's Office for Civil Rights—is directly relevant to the memo's fairness and non-discrimination principle.

Fairness and Non-Discrimination

Human Decision-Making is Biased

We applaud the draft memo's recognition that "AI applications have the potential of reducing present-day discrimination caused by human subjectivity" and its instruction that agencies consider whether AI applications "may reduce levels of … discrimination as compared to existing processes."

This is an important recognition because the discussion of bias in AI systems often ignores the fact that the alternative is to rely on human decision-making, which is typically more biased than AI systems. Humans come to a task with a lifetime of prejudices—many irrational or even malevolent—while a neural network, the core technology in most machine learning, starts out free of biases. Everything it knows, including potential biases, is learned mathematically from the data set of examples it is given. No malevolence is possible.

Moreover, in an AI system making decisions about humans, characteristics such as race and gender can be hidden from the system, making bias more difficult. Hiding such characteristics from a human is more problematic.

While the reasons why AI systems reach particular decisions can be difficult to determine, thus complicating the ferreting out of discrimination, the same is true for humans, who often don't fully understand why they made a particular decision. Moreover, humans will likely try to conceal any discriminatory reasons for their decisions,

Existing anti-discrimination laws can do much of the work

We applaud the draft memo's recognition that 1) "many AI applications do not necessarily raise novel issues," 2) "the broader legal environment already applies to AI applications," and 3) existing laws and regulations may be sufficient to handle the risks associated with AI. These points are certainly true for the problem of bias in AI systems because the United States has an extensive, well-developed body of anti-discrimination law.

It is important to remember that the AI systems currently in use are tools selected by and used by humans. Not every tool requires a new regulation. If a tool—say one used to aid hiring decisions—results in unacceptable discrimination, should it matter whether it's an AI-based tool or a less intelligent tool? Most importantly, the people responsible for selecting biased tools can be and often are held legally responsible

The draft memo recognizes that "the application of existing law to questions of responsibility and liability for decisions made by AI could be unclear in some instances." Future court rulings in discrimination cases will undoubtedly go a long way to make the answers to those questions clear. In addition, it may make sense to update existing discrimination laws and regulations to deal with any novel issues raised by AI-based tools.

In any case, applying existing discrimination laws to AI tools, perhaps with some tweaking, is much faster than promulgating a new regulatory regime to deal with bias in AI. Speed is particularly important here because of the rapid pace of technological development in AI.

There may come a time when AI systems are autonomous decision makers independent of human control. If so, a new regulatory regime will almost certainly be needed. But for now, such "strong AI" is "beyond the scope of this [draft] Memorandum."

Discrimination needs to be defined

Well-defined, consistent definitions of "fairness" and "non-discrimination" will need to be fleshed out. Though a relatively high-level document like the draft memo is probably not the appropriate place to do so, a process should be put in place to do so as the various agencies promulgate AI regulations.

The need for good definitions is particularly acute for AI systems that make decisions about humans because the very purpose of such systems is to discriminate. An AI system used to help make lending decisions must discriminate between people who are good and bad credit risks. An AI system used to aid hiring decisions must discriminate between people who are likely to be good employees and those who are not.

Any good definition of discrimination must start with a list of the types of discriminatory effects that are impermissible. That is not so easy to do, because AI systems don't discriminate in an intentional or malevolent way. Instead, a machine learning model looks for and relies on whatever correlations between variables can be used to improve the average accuracy of its output—whether the output is a prediction of credit worthiness or the identification of a face.

Because any alleged bias or disparate impact in the model's performance is the result of correlations that improved the model's accuracy, there are tradeoffs involved that must be taken into account by any definition of unacceptable bias. For example, AI systems are used in various jurisdictions to inform the bail decisions made by courts. If the data used to train such a model indicates that men are more likely than women to commit a crime while out on bail, the model can improve its accuracy by taking that correlation into account. Yet the resulting model will be one that has a disparate impact on men.

If the correlation was caused by a data set that does not accurately reflect the real world, the problem of disparate impact can be addressed by improving the data set. However, if bad data isn't the problem, the harm caused by a disparate impact on men must be weighed against the superior accuracy of a model that can take all correlations into account, in accordance with the draft memo's instruction that agencies should determine "which risks are acceptable and which risks present the possibility of unacceptable harm, or harm that has expected costs greater than expected benefits."

Disclosure and Transparency

We applaud the inclusion of "Disclosure and Transparency" as a guiding principle in the draft memo and agree with its observation that "transparency and disclosure can increase public trust and confidence in AI applications."

The bias issue discussed in the previous section is just one example of the importance of transparency. There are no easy answers about when disparate impacts should be acceptable in AI systems or how they can be eliminated. However, transparency can aid in answering such questions by allowing us to better understand the nature and source of such bias, as well as in ameliorating the damage to public trust that can result from bias, whether real or perceived.

As with the bias issue, the devil is in the details about what is meant by "disclosure" and "transparency." Again, a relatively high-level memo is probably not the appropriate place for defining those terms, no less discussing the various forms that disclosure and transparency can take and their relative benefits. However, those terms will have to be better defined before regulations referencing them can be issued.

We caution against the more extreme form of mandated transparency—typically including public disclosure of the weights between the artificial neurons in a neural network—that some who have written about the issue favor. As Curt Levey and Ryan Hagemann explain, "such disclosure will not tell you much, because the machine's 'thought process' is not explicitly described in the weights, computer code or anywhere else. Instead, it is subtly encoded in the interplay between the weights and the neural network's architecture."[1]

Moreover, they explain that such disclosure "may be harmful. Requiring disclosure of the inner workings of artificial-intelligence models could allow people to rig the system. It could also reveal trade secrets and otherwise harm the competitive advantage of a system's developers. The situation becomes even more complicated when sensitive or confidential data is involved."[2] Such harms are likely what the draft memo has in mind when it instructs that "What constitutes appropriate disclosure and transparency is context-specific, depending on assessments of potential harms [and] the magnitude of those harms," among other things.

We favor an approach to transparency, sometimes called "accountability," that is more flexible and offers a lighter touch. Levey and Hagemann describe this approach:

> *"[A]ccountability should include explainability, confidence measures, procedural regularity, and responsibility. Explainability ensures that nontechnical reasons can be given for why an artificial-intelligence model reached a particular decision. Confidence measures communicate the certainty that a given decision is accurate. Procedural regularity means the artificial-intelligence system's decision-making process is applied in the same manner every time. And*

[1] Levey, Curt and Hagemann, Ryan, *Algorithms With Minds of Their Own*, Wall Street Journal at A15, November 13, 2017.

[2] *Id.*

> *responsibility ensures individuals have easily accessible avenues for disputing decisions that adversely affect them."*[3]

A flexible, accountability-based form of transparency that doesn't mandate precisely what details need to be disclosed also makes the most sense in light of the draft memo's warning that "Rigid, design-based regulations that attempt to prescribe the technical specifications of AI applications will in most cases be impractical and ineffective, given the anticipated pace with which AI will evolve."

Sincerely,

Curt Levey
President
The Committee for Justice

[3] *Id.*