January 31, 2019

European Union
High-Level Expert Group on Artificial Intelligence
Ref: Stakeholders' Consultation on Draft AI Ethics Guidelines

The Software & Information Industry Association (SIIA) appreciates the opportunity to comment on the draft ethics Artificial Intelligence (AI) guidelines. SIIA supports the discussion of such guidelines with the caveat that guidelines will not be uniformly applicable to all AI applications given that AI has such domain-specific applications. Defense, health, autonomous vehicles, marketing/advisor bots etc. each pose their own unique requirements. Even more broadly, SIIA considers that there should be a global alignment on a definition for AI developed with public and private sector stakeholders both to assist public policymakers and the private sector. Furthermore, SIIA notes there is a discussion about possible regulation of AI in the EU. SIIA reiterates that given how quickly technology develops in unanticipated ways, it is crucial for regulation not to focus on emerging technologies, i.e. regulation should be technology, a precept for which there is wide international support. Instead, regulations should be designed to prevent harm to consumers and businesses and crafted to address domain-specific situations, rather than how AI could be used in general.

About SIIA

The Software & Information Industry Association (SIIA) is the principal trade association for the software and digital information industries. The more than 800 software companies, data and analytics firms, information service companies, and digital publishers that make up our membership serve nearly every segment of society including business, education, government, healthcare and consumers. As leaders in the global market for software and information products and services, they are drivers of innovation and economic strength – software alone contributes $425 billion to the U.S. economy and directly employs 2.5 million workers and supports millions of other jobs. For more information, please visit the SIIA Policy Home Page at www.siia.net.

Introduction

On September 17, 2017, SIIA released an Issue Brief entitled: "Ethical Principles for Artificial Intelligence and Data Analytics." [1] The draft AI guidelines are consistent in many ways with what SIIA says in the Issue Brief. Our comments provide additional information on how disparate impact analysis studies could be conducted. This information is likely most pertinent to the profiling and law enforcement use case mentioned on page 28 of the "Working Document for stakeholders' consultation" and the Non-discrimination point on page 25 of the consultation document. Furthermore, SIIA concurs with the relevance of the ten elements described as "Requirements of Trustworthy AI" and offers additional comments on the Robustness (8) and Transparency (10) elements.

[1] SIIA Issue Brief, "Ethical Principles for Artificial Intelligence and Data Analytics," September 15, 2017 http://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2017-11-06-160346-990

Non-discrimination - Conduct Disparate Impact Analysis to Check for Bias

With respect to the Expert Group's correct point in the Assessment List asking whether there "are processes in place to continuously test for such biases during development and usage of the system," SIIA considers that the way to address the possibility of bias is to conduct disparate impact tests as appropriate. Note: in this context, "disparate impact" means an impact that has a disproportionate adverse effect on vulnerable populations. The principles guiding disparate impact tests reflect the widespread international norm that high-stakes decisions about people should not disadvantage vulnerable populations based on characteristics such as their race, gender, ethnicity, or religion. See the italicized text below from the SIIA Issue Brief for when and how to conduct disparate impact assessments.

Since disparate impact occurs inadvertently, the only way an organization will discover on its own that its data practices have a disparate impact is to look for it. As noted above in the scope principle, organizations should put in place procedures and standards to determine when to conduct a full disparate impact assessment when they regularly develop, implement or use data analytic systems that might have a discriminatory effect on vulnerable groups. The following principles specify when a data analytic system should be subjected to a full disparate impact and what the elements of a disparate impact assessment are.

- *Organizations should evaluate a data analytic system for disparate impact when the design, implementation or use of that data analytic system has a significant potential for substantial and consequential discriminatory effects on vulnerable groups.*

- *A disparate impact assessment determines whether a data analytic system has a substantial disproportionate adverse impact on a vulnerable group, examines whether the use of the system advances legitimate organizational objectives and compares it to alternative systems that might have a lesser disparate impact.*

Organizations regularly operating in areas that have consequential impacts on people's lives should evaluate data analytic system techniques for disparate impact when the design, implementation or use of data analytic systems has a significant potential for discriminatory effects.

A disparate impact assessment has three steps. The first is to determine whether the data analytics system under review has a disproportionate adverse impact on a vulnerable group. This can be measured by standard statistical characteristics of the data analytic system such as departures from statistical parity or equal group error rates. Organizations should devise or adopt – in collaboration with academics, advocates, and independent technical experts – accurate and reliable guidelines and methodologies for detecting disparate impacts.

The second step is examination of how the data system in question serves organizational objectives. Notwithstanding any disproportionate adverse effect on vulnerable groups, a data analytic system can pass a disparate impact assessment if it furthers a legitimate organizational interest. Avoiding disparate impact cannot be a requirement to abandon the values and goals that constitute an organizations mission. But furthering a legitimate objective is not sufficient to pass a disparate impact assessment, because there might be an alternative system that also furthers organizational objectives, but does so with a smaller impact on the vulnerable group.

So, the third step in a disparate impact assessment is a comparison of the data system to alternatives. This step should involve an active search for alternatives to or modifications of the system being reviewed. It should not be restricted to an assessment of obvious or readily available alternatives. Organizations should develop and assess alternatives to algorithms with a disparate impact to ascertain the extent to which they achieve organizational objectives.

A data analytic system passes a disparate impact test, despite having a disproportionate adverse impact on a vulnerable group, when after an appropriate search for alternatives, an organization finds there is no alternative algorithm that furthers institutional objectives with a lesser impact.

Disparate impact assessments should be conducted at the same frequency as other reviews needed to ensure the validity and reliability of models. Especially in the case of advanced analytic systems that improve in use, impact assessments need to be conducted frequently.

It is crucial to emphasize the point above that the mere presence of a statistical disproportion involving protected classes is in no way a proof of legal liability for violation of non-discrimination laws. As noted above, these discrepancies are often an essential element in the use of algorithms to achieve legitimate business purposes. But they are an indication that further assessment is needed to determine the legitimate business interest served and whether there are alternative algorithms that could achieve the same result with less impact on the protected classes. For more detail on disparate impact assessments, see SIIA's Issue Brief on Algorithmic Fairness. [2]

Transparency - Communicate Key Factors in Scores and Evidence of Validity of Predictive Models

SIIA notes that that the Assessment List's points 8 and 10 do not mandate disclosure of source code of proprietary algorithms, and SIIA considers this outcome correct. Companies need to be able to choose proprietary business models (or not) as they develop algorithms. Moreover, disclosure of such source code could allow bad actors to game analytical systems that defeat their purpose, like for instance criminals intent on credit card fraud. For SIIA's view on Transparency and Explanations, see the italicized text below from the Issue Brief.

A key aspect of ethical use of data is an organization's willingness to be accountable to outside oversight about the processes and outcomes of data analytic systems. Accountability cannot be effective without transparency to the outside world and a commitment to conveying clearly and comprehensively how an organization's processes and standards address the ethical issues raised by data use, including how an organization assesses and remedies disparate impacts. Several U.S. and European regulations, described in the appendix on additional material, call for disclosures of explanations. The following principles regulate how an organization should approach these transparency questions.

• *Organizations should disclose what data they collect, the purposes for which it is used, and which analytic techniques and models are used to process data and produce an outcome.*

• *Organizations should provide explanations of how advanced modeling techniques produce their results, including disclosing, where available and appropriate, the key factors that contribute to the outcome of an analytic process.*

[2] SIIA Issue Brief, Algorithmic Fairness, September 22, 2016, http://www.siia.net/Portals/0/pdf/Policy/Algorithmic%20Fairness%20Issue%20Brief.pdf

• Organizations should publicly describe the model governance programs they have in place to detect and remedy any possible discriminatory effects of the data and models they use, including the standards they use to determine whether and how to modify algorithms to be fairer.

Trust in the fairness of a data analytic system relies on public awareness of data and the analytical systems used as well as the basis for organizational steps to detect and mitigate disparate impacts. Transparency about the process and standards used is especially important for disparate impact assessments, where ethical intuitions differ and social consensus on the right course of action might not be possible. The need to consult with public officials and the affected communities is especially strong in the cases, discussed below, of using sensitive variable in data analytic systems and determining how to navigate the tradeoff between accuracy and fairness when a data analytics system might not be able to fully satisfy both values.

Organizations do not need to disclose source code of proprietary algorithms for several reasons. Disclosure is not useful for accountability purposes, especially in the case of advanced analytical techniques that improve themselves in use. Source code disclosure would likely produce counterproductive efforts to game analytical systems in ways that defeat their purpose. Disclosure would allow anyone to use or benefit from systems that require extensive development resources, thereby weakening the economic incentive in creating these systems. For these reasons, disclosure has not been required for heavily regulated traditional scoring systems such as credit scores that have been in use for decades.

If organizations do not reveal their source code, they must take other steps to provide for transparency and accountability. Organizations should be prepared to communicate to outside parties the key factors that go into their scores, and to provide evidence on a regular basis of the continuing validity and reliability of the predictive models they use. Public trust in the fairness of algorithms requires sufficient disclosure so that people feel able to comprehend and assess the process used to produce insights that might have important effects on their lives.

Need for Sector Specific Guidelines

Regarding the trustworthy requirement, the draft guidelines say "…in different application domains and industries, the specific context needs to be taken into account for further handling thereof…" They also specify that:

> "While the Guidelines' scope covers AI applications in general, it should be borne in mind that different situations raise different challenges. AI systems recommending songs to citizens do not raise the same sensitivities as AI systems recommending a critical medical treatment. Likewise, different opportunities and challenges arise from AI systems used in the context of business-to-consumer, business-to-business or public-to-citizen relationships, or – more generally – in different sectors or use cases. It is, therefore, explicitly acknowledged that a tailored approach is needed given AI's context-specificity."

This is appropriate. The point the AI Study Group makes about regulation applies to ethics as well:

> "…attempts to regulate "AI" in general would be misguided, since there is no clear definition of AI (it isn't any one thing), and the risks and considerations are very different in different domains. Instead, policymakers should recognize that to varying degrees and over time, various industries will need

distinct, appropriate, regulations that touch on software built using AI or incorporating AI in some way.[3]"

The draft guidelines suggest that the final guidelines will emphasize this context-dependence by providing more specific guidelines for four distinct sectors. It might become clear in the discussion of these cases that the general guidelines are just elements to consider for appropriateness in a context, rather than requirements that must be implemented in all contexts.

SIIA recommends that this point be articulated more clearly and completely in the final version of the guidelines.

Relationship to Older Analytic Techniques

As the guidelines make clear and others have as well, AI techniques, and particularly, machine learning programs are different and perhaps better ways of accomplishing the same tasks that earlier analytical techniques attempted to achieve. For instance, a machine learning credit score might do a better job of detecting when a person is a good credit risk than one based upon standard logistic regression techniques, but they are both attempting to do the same thing. Similar remarks apply to machine learning programs aimed at assessing the risk of recidivism, AI-powered data programs designed to improve the delivery of public services, content moderation algorithms, and facial recognition programs. The key thing is not the statistical technique used but the risks and challenges presented by the attempt to accomplish these tasks through data and data analysis.

SIIA recommends that that the guidelines make it clear that the same ethical guidelines and regulatory rules apply to the application of analytics to achieve the same business or social objectives, regardless of the statistical techniques used.

On behalf of SIIA, I would like to thank you for the opportunity to comment. Please do not hesitate to contact us if you believe we can be of further assistance.

Sincerely,

Carl Schonander
Senior Director, International Public Policy
Software & Information Industry Association (SIIA)
1090 Vermont Avenue, NW
Washington, D.C. 20005
United States

[3] Artificial Intelligence And Life In 2030, One Hundred Year Study On Artificial Intelligence, Report Of The 2015 Study Panel, Stanford University, September 2016, available at https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf.