Contribution to a European Agenda for AI: Improving Risk Management, Building Strong Governance, Accelerating Education and Research

A Response to the European Commission White Paper "On Artificial Intelligence — A European approach to excellence and trust"

Baobao Zhang Follow
Jun 26, 2020 · 21 min read

By Adrien Abecassis, Justin B. Bullock, Johannes Himmelreich, Valerie M. Hudson, Jack Loveridge, and Baobao Zhang

Around the globe, governments draft policies to regulate artificial intelligence (AI). On February 19, the European Commission presented the EU's AI strategy with a new white paper. We submitted the following comments as a part of a public consultation and publish them here to continue the discussion.

Overview

As an interdisciplinary group of engaged scholars and policy analysts, the Governance of AI Research Group would like to respond to the EU Consultation for the EU AI White paper. In our responses below, **we address four guiding areas where we believe the EU Commission can build upon its current framework for addressing AI**. We offer general guidance in two underdeveloped, but vitally important, directions for the successful governance of AI, and we echo and expand two directions from the AI White Paper.

1. We build and offer recommendations for **creating a more comprehensive and holistic approach to risk management**.
2. We offer concrete guidance on **building out governance institutions** in a principles-based manner to appropriately empower these institutions to oversee the regulatory AI ecosystem.
3. We strongly support **increasing AI literacy throughout society** and give recommendations on how to do so.
4. We endorse **increasing funding for social sciences research** so that researchers can study the impact of AI on society to improve governance.

Recommendations for Improving the Risk-based Approach

The commission proposes a risk-based approach to guide AI regulation and outlines two cumulative criteria to determine whether an AI application is high-risk or low-risk. One of these criteria is whether an AI application is deployed in a sector that involves high risk.

We applaud approaching AI governance carefully and systematically through transparent and clear criteria as those provided by the proposed risk-assessment framework. We suggest five ways in which this risk-assessment framework can be extended or supplemented.

1. ***Going beyond the low-risk vs. high-risk binary, regulators should also consider the probability and severity of the risk, the subjects who could be impacted by the risk, and the type of risk.*** The AI White Paper proposes a sector-based binary classification of AI tools as either high-risk or low-risk. We suggest three ways of enriching this risk framework to provide more detailed information.

a. **Probability and severity.** Risk is generally defined as the probability and severity of the occurrence of an event.[1] This definition encourages rigorous and precise modeling as a best practice. This definition of risk is moreover consistent with EU general risk assessment methodology (COM(2013)76).[2] The notion of "risk" in the AI White Paper can hence be clarified in relation

to this definition, or this definition of risk (as a combination of probability and severity) could be adopted in a future version of the risk-assessment framework.

b. **Subjects at risk.** The risk-assessment framework could be extended by a dimension that **identifies potential or expected subjects at risk (i.e., which *individuals* or *groups* are exposed to the risk)**. A given risk will differ in how it affects different people or groups (e.g., gender, ethnicity, age group), and "subjects at risk" is hence relevant information. Including a dimension of "subjects at risk" in the risk assessment framework, and understanding "subjects" as including groups, not merely individuals, is supported by three specific considerations.

First, building trustworthy AI systems involves building software that's ethical and socially acceptable. Identifying the subjects at a given risk helps anticipate and proactively mitigate ethical and legal problems of algorithmic discrimination and unfairness. These problems of fairness and discrimination are significant roadblocks to social acceptability that stand against the widespread uptake of AI.[3] An AI policy framework, therefore, should be sensitive to such social acceptability obstacles and address them by identifying subjects who are exposed to risks.

Second, identifying potential subjects of risks provides a connection to social science research and participatory development. Once subjects of risks are identified, input from these affected groups and their representatives can be made to count in policymaking and AI development.[4] Including "subjects at risk" as part of the risk-assessment framework builds a bridge to our recommendation concerning the importance of social science research for the systematic study of potential harm (see "Increase Funding for Social Science Research" below).

Third, identifying subjects of risks enables organizing by civil society, collective action by citizens, and strengthens citizens' voices. Because affected individuals and groups are identified this provides information to affected stakeholders and allows for collective actions and policy innovations on behalf of these stakeholders.

Instead, the risk assessment framework must **take the risk for society into account**.[5] Other regulators do this routinely: health authorities, for example, assess the risk of an antibiotic not only based on the risk/benefit balance at the individual level but also by integrating collective risks (e.g., increasing the resistance of bacteria in the ecosystem). Similar mechanisms must be explicitly provided for in AI risk assessment.[6] This is because AI tools can similarly contribute to collective harm, for example when

algorithmically curated social media create filter bubbles that exacerbate polarization.

Moreover, it has been shown that algorithms lead to collective biases that might reinforce the polarization of societies. Algorithms for predictive policing, loan granting decisions, or hiring tend to build statistical groups over iterations and reinforce the differences between these groups (see also our recommendation on 'type of risk' below).[7] [8] Thus, **the "subjects at risk" should include collective risk subjects such as groups.**

c. **Type of risk.** The risk-assessment framework should be supplemented with information about the type of risk. Different **types of risk include health risks**, **privacy risks, and opportunity risks** (e.g., unfair dissemination of information about scholarships or job listings).[9] Extending the risk-assessment framework with such identification of the type of risk invites rigor of risk analysis, offers a further bridge to social science research, and allows for connections to existing ethical and legal frameworks (such as human rights conventions).

2. ***The framework should consider the upsides as well as the downsides of AI tools.*** Risk is often a negative notion associated with potential harms or injuries. As with any risk-based framework, the proposed guidance is purely defensive by identifying downsides. Research and practice shows that there can be various benefits to AI tools that improve human decision making.[10] [11] [12] [13] Given the existing criteria for AI that the commission has proposed, the framework could be supplemented with measures to operationalize the extent to which aspirational measures are met by an AI application (e.g. transparency, fairness, etc). Ideally, a policy to govern AI does not only include a framework to measure the downside risks but also a framework to assess the potential benefits,[14] such as improvements to fairness, health, privacy, equity or efficiency.[15]

3. ***The framework should assess the risk of tasks in addition to sectors.*** The framework proposes to assess risks based on the industry sector. We suggest using "tasks" as an additional basis for evaluating risk.[16] Sectors differ greatly internally with respect to the risk that AI tools pose. For example, the health care sector appears to exhibit greater risks than municipal garbage collection, but tasks within the latter could be high risk. As municipal garbage collection transitions to autonomous vehicle technology, very mundane driving decisions, such as whether the vehicles should avoid left turns, can have a significant negative impact on population safety.[17] Likewise, accounting may as a whole appear to be a low-risk sector, but individual tasks and practices that individual accountants engage in might carry significant risks. Hence, concentrating on tasks within sectors allows for a more fine-grained risk analysis.

4. ***The framework should be adapted to consider risks from general-purpose AI systems.*** As AI research progresses, novel AI systems will be increasingly general-purpose. By "general-purpose AI", we understand an AI system that can be deployed for more than one task. One example of that is the text generation model known as GPT-3.[18] A text generation model can be used to create a customer service chatbot but also generate fake news articles.[19] General-purpose AI systems hence are able to complete different kinds of tasks and could be deployed in different sectors. As AI research and innovation progresses, we expect that the number and proportion of general-purpose AI tools to increase. Therefore, policymakers should test whether the current risk-assessment framework could be meaningfully applied to general-purpose AI. If not, they should update the risk-assessment framework to accommodate general-purpose AI.

5. ***Risk should be anticipated already in the development phase.*** We argue that the importance of responsible innovation should be emphasized as part of a risk-assessment framework. The probability and severity of risk that eventually results from an AI application can be anticipated and mitigated during the design and development of the AI application. The risk of an AI application is, at least to some limited extent, under the control of the developer. This aspect of technological development and the importance of responsible innovation could be recognized more prominently in AI regulation and, perhaps, even be reflected in a possible extension of the risk-based framework consistent with other regulations and guidelines.[20]

Guidance on Building Governance Institutions

We would like to begin this section with the reminder that the discussion on the different levels of regulation and governance could be clarified in the paper. Algorithm regulation and governance could be carried out at **three important points** in the algorithmic process, from upstream to downstream:

1. Regulating the data upstream: this involves setting rules and principles on who has access to data, who has the right to use data, under what conditions, with what consent or authorization.

2. Regulating the algorithms themselves to prevent harmful effects: auditability, certifications, in some cases full transparency of the source code and data used.

3. Regulating the social impact of applications downstream: setting rules and principles of what governments and private actors should or should not produce, up to developers to elaborate the technical solutions to comply with them.

Regulation at each point in this process is necessary but by itself incomplete. Effective regulation will need a combination of approaches that vary depending upon the domains, applications, and dimensions of risk as stated in the previous section. The Commission should clarify the levels which it wishes to prioritize and how this relates to other existing regulations and international declarations and commitments to which the EU has subscribed. Policymakers need to consider how the AI governance framework outlined in the White Paper relates both to the General Data Protection Regulation and the European Data Strategy.

With each of these points of regulation and governance in mind, we describe some "principles of enforcement" concerning AI applications (especially when used as decision-making systems), along with thoughts on mechanisms designed for that enforcement:

1. **Principles of Enforcement**

a. Everyone has the right to **know** when they are engaging with an AI system. They should be notified and be shown a standardized identification label with the contact information for the liable party (see below).[21]

b. Everyone has the right to **appeal** to a human being in the entity making the decision. So, for example, after knowing they were denied an interview by an AI hiring application, an individual would have the right to appeal to a human being in the hiring organization. This human point of appeal should be empowered to make a decision without recourse to the AI app under question. Only a human being can see when an algorithm has veered from its intended purpose.[22] Successful appeals should bring about a renewed audit of the algorithm in question.

c. Everyone has the right to **litigate** the harm caused by AI applications. The liability rests with the vendor of the algorithm who has sold the product to the entity whose deployment of the AI system caused harm. In some cases, that may be the same party as the vendor, but often it is not. End party entities that were negligent in their purchase from the vendor, having failed to check for due diligence on the part of the vendor with regard to appropriate auditing, may also be held partially liable.[23]

2. **Harm Avoidance Mechanisms:** The rights to **know**, to **appeal**, and to **litigate** will require mechanisms to make their implementation possible:

a. Deployment of AI applications will require capabilities to ensure:

- Mandatory notification is given to all who encounter the system, accompanied by

- Mandatory labeling identifying who was the vendor who sold the system to the decision making entity deploying the system.

b. Within each decision making entity that uses AI applications, a Human Appeals Department or its like should be established. Those who staff this department must understand the intentions of the decision-making process. In addition to being enabled to overturn or revise AI decisions, this department must also be empowered to demand changes to the algorithms be made by the vendors where the algorithms have been identified as veering from the original intent of the customer. In the case the decision-making entity and vendor are the same, this Human Appeals Department would be empowered to demand a renewed audit and changes to the existing algorithm.

c. Due diligence on the part of customers of these vendors of AI applications demands that there be a certification process for algorithms before they are sold or deployed; independent external auditors must certify that, to the best of their knowledge, a particular algorithm does not cause harm. "Harm" in this case is defined not only at the individual level, but also at the collective level, and pertains to tangible harm such as job loss or physical injury, as well as less tangible harm, such as the reification of bias against protected classes within a society. All source code, as well as training data, testing data, logs, and other pertinent material, must be made available to independent external auditors for such certification. Changes to the algorithm over time must also be submitted for certification. Should an entity purchase a system without such certification, the entity would have demonstrated negligence and thus would become equally liable with the vendor for any harm caused.

d. In litigation, plaintiffs asserting an AI system has caused them harm are entitled to access the source code and any other relevant material of the algorithm under question, and vendors must waive any rights to secrecy or opacity regarding the code.

e. The government must have the power to prevent or reverse the deployment of a system where it cannot be shown that risks have been satisfactorily mitigated.

f. The retirement of an AI tools system must ensure that claims can continue to be litigated. This means that source code, training data, and testing data must still be retained by the vendor even if the vendors retire the system. Retention may also be accomplished through the certification companies or independent archives established for that purpose, as well.

3. **What Would Be Required to Establish These Mechanisms**

a. The first need is to stand up a cadre of trained independent algorithm auditors. This will require an accelerated educational effort and the establishment of independent companies whose certifications will carry weight because they are impartial and independent.

b. The second need is to develop the capabilities of lawyers and legal specialists to litigate algorithm liability cases. This, too, will require an educational effort by law schools.

c. A new track within the Human Resources field will need to be created that focuses on the training of new HR Human Appeals Officers for AI tools.

d. Regulations governing the new mechanisms will require the creation of public institutions such as agencies, departments, and commissions, as well as the enumeration of their powers to prevent or reverse the deployment of AI tools.

e. Education of the public concerning their rights to know, to appeal, and to litigate must be promulgated. Citizens should ideally understand in broad terms how AI systems function.

We provide further detail concerning these requirements in the following section.

Increase AI Literacy Skills Within the Public Sector, the Private Sector, and the General Public to Enable Robust Auditing and Appeals Capability

To successfully regulate AI applications, the EU will need to invest significantly in human capital by training auditors, regulators, and public sector employees, as well as educating the public. Building AI literacy is essential for those making decisions about the funding, procurement, and deployment of AI systems.[24] Algorithm auditors are needed to evaluate the impacts of machine learning algorithms; they should be competent in addressing the following questions:

- What are the costs and benefits of deploying the AI application? In particular, what is the level of risk to individuals and groups if a given AI application is deployed?
- How has the AI application been tested for fairness, robustness, and safety, not only with regard to individuals but also with groups and society at large?
- What risks can be predicted to arise when the AI application interacts with other systems? For example, credit score AI systems may interact

with the loan application AI systems to produce unintended negative consequences.

The AI White Paper encourages the adoption of AI applications by the public sector. Therefore, public sector employees should be trained to evaluate AI applications before procurement. Those running public agencies need to understand the various risk and technical aspects associated with the use of AI applications. Managers within these agencies also need to understand how to design information flow systems with AI applications in mind. They need to understand the risk-based decision-making strategies illuminated here and throughout the literature. Even so, the government will not have the resources to do the job alone.

Therefore, we identify three specific high need areas for human capital development; these include:

1. ***Recruit and train independent algorithm auditors.*** The auditors should be trained by entities independent of vendor tech companies, such as universities. We envision that for a healthy AI ecosystem to exist, independent and impartial auditing firms that check the claims made by tech companies will be necessary in the private sector.[25] The government will not have the resources to perform all the audits required. But insofar as the government must have its own capability in auditing, as well, these accreditation programs should be made available to public sector employees.
2. ***Develop the capabilities of lawyers and legal professionals to litigate algorithm liability cases.*** Law schools should offer courses and training in AI and the law. Lawyers and legal professionals will also benefit from basic education in the auditing of algorithms since many expert witnesses in these cases will be professional algorithmic auditors.
3. ***Create a new track within the Human Resources (HR) field that focuses on the training of new HR Human Appeals Officers for deployed AI systems.*** The HR field already trains professionals in relevant skills, such as grievance mediation, identification of discriminatory differential treatment, legally compliant decision making with regard to hiring, and other direct effect processes, among many others. Training HR professionals to handle legally mandated human appeals processes for AI systems is a natural extension of this field of expertise.

Finally, the general public, the end-users of so many of these AI applications, need to be better educated on the basics of how AI systems work and how they impact their lives. Increasing awareness can be done through a widespread public service announcement (PSA) campaign. These

PSAs could be delivered in partnership with other like-minded institutions and should emphasize the public's right to know, to appeal, and to litigate. Members of the general public must never feel helpless or ignorant in the face of an increasing number of AI systems being deployed.

Increase Funding for Social Science Research

One key element in training and developing the needed human capital for society is increasing funding for social science research. Social science research can inform policymakers and citizens on how to use AI applications in ways beneficial to society. Additionally, social scientists could counterbalance the influences of AI firms by identifying the flaws in AI systems. To this end we provide the following comments:

1. ***Social science research has uncovered major flaws in automated systems.*** Information studies scholar Safiya Noble uncovered racial and gender bias in search engine algorithms.[26] Gender studies and political science scholar Virginia Eubanks documented how errors in algorithms used in welfare decision-making, housing allocation for homeless people, and preventative child protection interventions further harm marginalized communities.[27] Economist Sendhil Mullainathan, working with public health researchers, detected anti-Black racial bias in a widely-used health care algorithm.[28] In their social science-informed work, computer scientists Joy Buolamwini and Timnit Gebru found that commercial facial recognition systems are biased against Black people, particularly Black women.[29] These examples show how central social science research is to exposing serious harms that could be caused by AI systems.
2. ***The gap between research in AI and relevant work across the social sciences is growing and risks obstructing opportunities for productive collaboration.*** Dashun Wang and his research team at Northwestern University's Kellogg School of Management suggests that developments in AI research and the social sciences have not kept pace with one another.[30] Wang et al.'s study implies that AI-specific researchers are increasingly publishing their work within topic-specific forums while AI research is notably absent from references made by social scientists in their own work. This suggests that AI research is becoming isolated from the sociologists, economists, and philosophers who can best inform and benefit from developments within the field. Wang cautions that this growing gap will hinder the development of AI-driven technologies, arguing for increased collaborations between AI researchers and the broader social sciences that create a two-way street of exchange. Other recent work, notably by Tim Miller, demonstrates ways in which the social sciences prove uniquely capable of sharpening and improving the

capacity of AI technologies while also improving the study of the ethical concerns related to those same technologies.[31]

3. ***Investments in areas of exchange and overlap between AI and the social sciences would support the broader goals of the European Commission surrounding innovation and oversight.*** With the European Commission increasing its annual investments in AI by 70% through the Horizon 2020 research and innovation program, there is increasing support for AI research centers across Europe.[32] As part of this effort, a corresponding investment is needed to enable existing social science research centers to develop research agendas that productively engage developments in AI. At the level of academic research, the social sciences are well-positioned to provide the insights and scrutiny needed to refine the ethical application of AI within society. Just as the AI-driven technologies can benefit from a clearer understanding of social, legal, and ethical challenges, AI applications can greatly enhance conventional social scientific research at a time when such research at institutions of higher learning across Europe is under increased financial strain and administrative scrutiny. Increased investments in applied and basic research related to AI applications across the social sciences would improve the technologies under development. At the same time, industry-independent research provides credible evidence for oversight of tech companies.[33]

4. ***Building upon existing capacity within Europe would enable the creation of a public-facing, policy-focused research environment at the forefront of developments within AI-driven technologies.*** Europe already possesses a robust independent technology assessment network driven by social scientists and higher education institutions in the form of the European Technology Assessment Group (ETAG), with the Institute for Technology Assessment and Systems Analysis (ITAS) at the Karlsruhe Institute of Technology (KIT) serving as its leading partner.[34] The ETAG provides scientific advisory services for the European Parliament across a broad spectrum of technology policy issues. Increased support should be directed toward ETAG to explicitly advance projects in partnership with European universities and research institutes that adopt a dual directional approach to the relationship between AI and the social sciences. ETAG is just one example of an existing research group dedicated to technology assessment with a specific capacity in attending to issues surrounding AI. Other networks might also be considered for this function, particularly drawing upon the capacity of European research centers and universities. Beyond applied research for the purposes of creating specific technology assessment products and investigating the impacts of AI technologies upon society, dedicated support is also needed for basic research into areas of overlap between AI and the social sciences. Such investment in

research on the governance and accountability of algorithms should be one of the priorities of the EU's next multiannual financial framework.

Conclusion

The Governance of AI Research Group would like to applaud the EU Commission's efforts in the direction of publishing the AI White Paper and providing a call for public consultation. **As a group of scholars, we are deeply concerned about the use of AI systems across many domains in the private and public sectors.** We see the power of AI applications and hope to work collaboratively with governments, companies, nonprofits, and other researchers to secure that AI applications serve human interests and do not violate human rights.

It is to this end, and in conjunction with what we perceive as overlap in our values with those of the EU Commission, that we have offered 4 general points for the Commission to consider as it seeks to improve the regulatory framework and governance structure around AI applications and systems.

First, we offer edits and improvements to the risk-based framework. While the members of the group generally endorse a risk-based approach, we note a number of deficiencies and provide numerous suggestions to make the framework more effective and overall reflective of both the opportunities and threats that AI applications present to society.

Second, we provide guidance on building governance institutions for AI governance. We argue that AI governance needs to take place across the inputs to the algorithm, the algorithm itself, and the use of the algorithm by individuals and institutions throughout society. We highlight specific principles that should be followed and propose institutions and the human capital needs to fill those institutions for the implementation and enforcement of regulatory changes. Carefully considering both the governance structure of algorithms and the needed governance institutions are crucial aspects to consider throughout this regulatory process.

Third, we argue for an increase in AI literacy skills within the public sector, the private sector, and the general public to enable robust auditing and appeals capability.

Fourth, and finally, we argue for an increase in social science research funding because the discipline plays a critical role in generating knowledge about how AI applications interact with society.

It is our hope that the arguments we provide here will help inform the EU Commission's AI White Paper consultation process and the growing policy

dialogue around AI governance. We encourage others to join in this conversation as well.

Statement Provided by the Governance of AI Research Group

Contributors include:

Adrien Abecassis, Fellow at the Project on Europe and the Transatlantic Relationship, Kennedy School of Government, *Harvard University*.

Justin Bullock, Associate Professor of Public Service and Administration, the Bush School of Government and Public Service, *Texas A&M University*.

Johannes Himmelreich, Assistant Professor of Public Administration and International Affairs and Senior Research Associate of the Campbell Public Affairs Institute, the Maxwell School of *Syracuse University*.

Valerie M. Hudson, University Distinguished Professor in the Department of International Affairs at the Bush School of Government and Public Service at *Texas A&M University*.

Jack Loveridge, Associate Fellow of the Weatherhead Center for International Affairs, *Harvard University*.

Baobao Zhang, Fellow at the Berkman Klein Center for Internet & Society, *Harvard University* and Research Affiliate of the Centre for the Governance of AI, *University of Oxford*.

References

1. Bullock, Justin B., Robert A. Greer, and Laurence J. O'Toole Jr. "Managing risks in public organizations: A conceptual foundation and research agenda." Perspectives on Public Management and Governance 2.1 (2019): 75–87. https://doi.org/10.1093/ppmgov/gvx016.
2. Available at https://ec.europa.eu/docsroom/documents/17107/attachments/1/translations/en/renditions/pdf
3. Young, Matthew M., Justin B. Bullock, and Jesse D. Lecy. "Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration." Perspectives on Public Management and Governance 2.4 (2019): 301–313. https://doi.org/10.1093/ppmgov/gvz014.
4. For an example of participatory policymaking, see Chung, Anna, Dennis Jen, Jasmine McNealy, Pardis Emami Naeni, and Stephanie Nguyen.

"Project Let's Talk Privacy." Technical Report, MIT Media Lab. 2020. https://letstalkprivacy.media.mit.edu/ltp-full-report.pdf

5. Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (*FAT* '19*). Association for Computing Machinery, New York, NY, USA, 59–68. DOI:https://doi.org/10.1145/3287560.3287598
6. This should also take into account that algorithms are evolutionary throughout their life (unlike drugs whose molecules do not change), requiring specific mechanisms ensuring, for instance, that the functionalities of an algorithm affecting democracy or justice do not change without the knowledge and the consent of the people affected by this change.
7. Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (*FAT* '20*). Association for Computing Machinery, New York, NY, USA, 525–534. DOI:https://doi.org/10.1145/3351095.3372878
8. Lum, Kristian, and William Isaac. "To Predict and Serve?" *Significance* 13, no. 5 (October 1, 2016): 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x.
9. For an alternative risk framework see https://ethicalos.org.
10. Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133, no. 1 (February 1, 2018): 237–93. https://doi.org/10.1093/qje/qjx032.
11. Abebe, Rediet, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. "Roles for Computing in Social Change." *ArXiv:1912.04883 [Cs]*, January 28, 2020. https://doi.org/10.1145/3351095.3372871.
12. Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366, no. 6464 (October 25, 2019): 447–53. https://doi.org/10.1126/science.aax2342.
13. For several practical examples of data science for social good, see: https://www.dssgfellowship.org/projects/
14. Bullock, Justin B., Robert A. Greer, and Laurence J. O'Toole Jr. "Managing risks in public organizations: A conceptual foundation and

research agenda." Perspectives on Public Management and Governance 2.1 (2019): 75–87. https://doi.org/10.1093/ppmgov/gvx016.

15. Matthew M Young, Justin B Bullock, Jesse D Lecy, Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration, Perspectives on Public Management and Governance, Volume 2, Issue 4, December 2019, Pages 301–313, https://doi.org/10.1093/ppmgov/gvz014
16. Bullock, Justin B. "Artificial intelligence, discretion, and bureaucracy." The American Review of Public Administration 49.7 (2019): 751–761. https://doi.org/10.1177/0275074019856123
17. Himmelreich, Johannes. "Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations." *Ethical Theory and Moral Practice* 21, no. 3 (May 17, 2018): 669–684. https://doi.org/10.1007/s10677-018-9896-4.
18. Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. "Language Models Are Few-Shot Learners." *ArXiv:2005.14165 [Cs]*, June 4, 2020. http://arxiv.org/abs/2005.14165.
19. Kreps, Sarah and Miles McCain. "Not Your Father's Bots: AI Is Making Fake News Look Real." *Foreign Affairs*. August 2, 2019. https://www.foreignaffairs.com/articles/2019-08-02/not-your-fathers-bots
20. http://www.businessofgovernment.org/sites/default/files/Risk%20Management%20in%20the%20AI%20Era.pdf
21. See, for example, Brian Higgins, "Recognizing Individual Rights: A Step Toward Regulating Artificial Intelligence Technologies," *Artificial Intelligence Technology and the Law*, 10 January 2018, http://aitechnologylaw.com/2018/01/recognizing-individual-rights-regulating-ai/
22. See, for example, Sherif Elsayed-Ali, "Why Embracing Human Rights Will Ensure AI Works for All," World Economic Forum, 13 April 2018, https://www.weforum.org/agenda/2018/04/why-embracing-human-rights-will-ensure-AI-works-for-all/
23. An excellent resource is Rashida Richardson, Jason M. Schultz, and Vincent M. Sutherland, "Litigating Algorithms," AI Now Institute, September 2019, https://ainowinstitute.org/litigatingalgorithms-2019-us.pdf
24. Horowitz, Michael and Lauren Kahn. "The AI Literacy Gap Hobbling American Officialdom." *War on the Rocks.* 14 January, 2020, https://warontherocks.com/2020/01/the-ai-literacy-gap-hobbling-american-officialdom/

25. Clark, Jack, and Gillian K. Hadfield. “Regulatory Markets for AI Safety.” arXiv preprint arXiv:2001.00078 (2019).

26. Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press (2018).

27. Eubanks, Virginia. A*utomating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* New York: St. Martin’s Press (2018).

28. Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. “Dissecting racial bias in an algorithm used to manage the health of populations.” Science 366, no. 6464 (2019): 447–453. https://science.sciencemag.org/content/366/6464/447

29. Buolamwini, Joy, and Timnit Gebru. “Gender Shades: Intersectional accuracy disparities in commercial gender classification.” In *Conference on Fairness, Accountability and Transparency,* pp. 77–91. 2018. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

30. Frank, Morgan R., Dashun Wang, Manuel Cebrian, and Iyad Rahwan. 2019. “The Evolution of Citation Graphs in Artificial Intelligence Research.” *Nature Machine Intelligence* 1: 79–85.

31. Miller, Tim. 2019. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” *Artificial Intelligence* 276: 1–38.

32. https://ec.europa.eu/digital-single-market/en/artificial-intelligence

33. Matias, Nathan. “Why We Need Industry-Independent Research on Tech & Society.” Technical Report, Citizens and Technology (CAT) Lab, Cornell University (January 2020). https://citizensandtech.org/2020/01/industry-independent-research/

34. https://www.itas.kit.edu/english/etag.php

AI Governance Ai Ethics Public Policy European Union

Learn more.

Medium is an open platform where 170 million readers come to find insightful and dynamic thinking. Here, expert and undiscovered voices alike dive into the heart of any topic and bring new ideas to the surface. Learn more

Make Medium yours.

Follow the writers, publications, and topics that matter to you, and you’ll see them on your homepage and in your inbox. Explore

Share your thinking.

If you have a story to tell, knowledge to share, or a perspective to offer — welcome home. It’s easy and free to post your thinking on any topic. Write on Medium

Medium

About Help Legal