Why Automatic AI Ethics Evaluations are Coming, and How They Will Work

James Brusseau and Giovana Meloni Craveiro

Abstract

Ethics evaluations of companies that function with AI at their core are increasingly required by regulation and law in Europe and the United States. Investors in AI-intensive companies also seek ethics evaluations as part of the nonfinancial information they gather about corporate performance, especially as it relates to privacy and algorithmic fairness. The result is an increasing demand for the evaluations. However, the costs and time necessary to perform an AI ethics audit are high, even prohibitive. To solve the problem, natural language processing and machine learning can be employed to automate the process. The proposal is that much of the work of AI evaluating can be accomplished more efficiently by machines than by humans. To show how automated ethics reporting may work, this paper describes a project currently underway at Pace University in New York, and the University of Trento in Italy. The project endeavors to apply artificial intelligence to the task of producing AI ethics evaluations.

1. AI ethics evaluations are coming

Companies that employ artificial intelligence (AI) at the core of their operations are coming under pressure to evaluate the technology in ethical terms. From the legal side, evaluations are increasingly required by governmental regulation. The European proposed AI Act from April 2021 follows up on the General Data

Protection Regulation (GDPR) and stipulates evaluations of AI applications that present physical or psychological risks to humans. These will include most significant applications in healthcare, hiring, insurance, and law enforcement.[1] Sometimes labelled as "AI audits" or "algorithmic impact assessments," and sometimes categorized as "post-market monitoring," the formal content of these evaluations remains a matter of discussion, but most proposals gravitate around the European Community's *Ethics Guidelines for Trustworthy AI*.[2] That document provides categories for measurement that include human autonomy and privacy, as well as algorithmic fairness, and also the explainability of algorithmic determinations. These principles can be understood as metrics, as ways of measuring how well an AI technology is functioning in ethical terms.

Moving from the European legal regime to the United States, regulatory legislation has been implemented at both the state and national level. In the area of privacy, the California Consumer Privacy Act shares significant goals and requirements with the European General Data Protection Regulation (GDPR). More broadly, the Algorithmic Accountability Act of 2022 is proposed legislation in Washington DC that requires companies to conduct assessments for bias, effectiveness, and similar factors when automated decision systems make "critical decisions" that have significant effects on people's lives.[3] The areas of education, employment, financial planning, and legal decision-making are early targets of the rules, and they would initially apply to data-intensive companies with more than $50 million annual revenue. The Algorithmic Justice and Online Platform Transparency Act of 2021 is complementary legislation also under consideration, and contributes to a wave of impending rules across Europe and North America that promise to "benefit the growing sector of AI auditing and monitoring services that could provide impact assessments."[4]

So, there are increasing legal and regulatory pressures for the creation of AI ethics audits. Additionally, contemporary economic forces encourage the ethical evaluation of AI-intensive companies. Under the heading of nonfinancial risk, investors are paying increasing attention to how the humanistic and social elements of a business can impact downstream profits. Research shows that high or favorable ratings tend to correlate positively with valuation and profitability, and negatively with volatility.[5] For that reason, capital allocation and subsequent investment benefits have surged around companies that can account positively for performance in terms of human values.[6]

At this writing, Facebook is involved with the European Court of Justice – the European Union's top court – in a standoff over a ruling commonly named *Schrems II*. It protects users' privacy to the degree of threatening the company's ability to function profitably. In response, the parent company, Meta, is considering the possibility of withdrawing Facebook and Instagram from European users. Regardless of how the conflict resolves, the impact of AI ethics remains as an example of how accounting for a company's ability to protect users' personal information does convert into information that helps investors avoid risk and increase returns.[7] On one side, foreseeing a company's vulnerability in terms of privacy may prevent poor investments, and on the other a company that scores poorly in terms of privacy may take corrective actions before financial damage occurs.

Tesla provides another example. For that company too, profit will be dialed up or down partially depending on the human experience of the cars' autopilot technology because, as a recent academic study has revealed, public sentiment, social media, and regulation entwine to determine manufacturer liability in the murky legal area of self-driving cars.[8] In terms of AI ethics, the question here is about how well the company scores in terms of safety, and the answer will depend on numerous factors, including the way safety is defined: is it measured as deaths per miles driven, or by some other metric? In any case, the larger result is that the materiality of an ethical evaluation on the financial level is evident at Tesla, and across the information technology industry.

The broad conclusion is that measuring ethics matters legally and financially. It will affect research, development, deployment, and profitability of AI technology.

2. What is the state of the art in AI ethics evaluations?

Evaluating the ethics of technology companies is very different from the ethics-rating of the industrial economy. Questions about factory labor conditions and toxic waste slip in terms of materiality, while digitally-oriented questions rise: debates about filter bubbles and privacy violations gain traction and attention as the fourth industrial revolution progresses. In terms of institutional documents, the seventeen categories listed as the United Nations Sustainable Development Goals (UN SDGs) decline in relevance in the sense that the metrics are extremely broad

and general (No poverty, zero hunger, and similar), and what increasingly matters are the tighter categories fitted to current technologies. Does the AI promote human autonomy, protect user privacy, make fair determinations? Instead of UN SDGs, the benchmark documents today include the European Commission's *Ethics Guidelines for Trustworthy AI,*[9] and the *Opinion of the German Data Ethics Commission*.[10]

Surveys of recent developments around AI ethics evaluations report that the number of participants in the field is growing rapidly to meet the increasing demand, and participants include both academic organizations and private enterprise, especially in the areas of law and consulting.[11] While the number of organizations offering the service is growing, the basic structures for performing the evaluations are converging, especially around the EC *Ethics Guidelines.*

An illustrative example of AI ethics auditing today as executed on the academic side emerged from the Frankfurt Big Data Lab in Germany in 2020. A PhD-level team of philosophers, computer scientists, doctors and lawyers united to approach AI-intensive startup companies in the field of medicine, and to collaboratively explore the technological development's ethical aspects.[12] To begin, the company's AI designers were invited to meet the data ethics team, and together the collaborators outlined a socio-technical scenario that staged the specific functioning of the artificial intelligence technology, and its interaction with users. From there, the group analyzed the interactions through the lens of AI ethics principles, and eventually concluded with case-study papers subsequently published in academic journals.[13] Going forward, the reports may be leveraged by policy experts and financial advisors to produce an understanding or a rating of the technology's humanist – as opposed to purely financial – profile.

The process succeeded. But, it equally revealed its own defectiveness: there is no way to implement the practice at scale. The limitation exists for two reasons. One is cost. While members of the Frankfurt-based group performed their ethics evaluations for scholarly reasons, professionalizing the process would immediately become prohibitively expensive. The group included medical doctors, lawyers, and computer scientists, and while it may be that philosophers and ethicists are plentiful and inexpensive at an hourly rate, the remaining members of the team would require significant financial incentives to continue the work beyond a few pilot projects dedicated to exploring the ethical terrain of AI.

Concretely, the group's first project involved a limited liability healthcare startup with a recent funding round of less than 4 million dollars, which implies that the cost of the ethics evaluation – had participants charged at market rates – would have amounted to a significant share of the value of the entire company.

The second obstacle is time. The evaluation process stretched over months, and then more months were required for the formal writing, submission, and publication of the results. Part of the reason for the slow pace was the voluntary nature of the effort, but the greater share of the explanation was the learning curve. For participants unfamiliar with the history and application of ethics, the language and reasoning can be daunting. Medical, legal, and statistical minds convert into ethically aligned thought only with difficulty, and entire meetings of the group collapsed under the weight of mutual misunderstanding,[14] and the subsequent hours required to find common grounds of ethical discussion.

So, procedures for ethically evaluating AI-intensive companies have been established and implemented in the real world. We know what to do. At scale, however, AI ethics audits as currently implemented are not feasible because they cost too much and take too long.

These two problems contain a larger threat. Already, studies are appearing to warn that under current conditions, the implementation of AI ethics requirements as stipulated by the European AI Act will prove catastrophic for innovation and ironically defeat themselves.[15] The financial costs and time required for compliance will catch technology companies in a regulatory paradox: "More regulation equals less regulating because innovation is stymied until there is nothing left to restrict."[16]

3. How can AI ethics evaluations be produced less expensively and more efficiently? Decentralization and acceleration

One way to answer the demand of regulators and investors for AI ethics evaluations that do not stymie innovation is to decentralize and accelerate the process.[17] Decentralization means the sources of evaluation expand beyond a tight core of highly paid experts to include diverse sources of publicly available information. So, instead of insular auditing organizations analyzing a few preferred data sources, vast public information is gathered, rendered

comprehensible and organized by artificial intelligence. And, acceleration does not just mean speeding the evaluative process, but also developing structures for evaluation that may be quickly transposed from one technology and company to another. So, instead of the grinding labor of specialized human evaluators trained in narrow domains of expertise, the scoring is accomplished by the workings of algorithms that are interchangeable, that apply as well to an established social media company as they do to a startup healthcare or financial enterprise.

One model for this decentralization and acceleration comes from AI-driven Environmental, Social and Governance (ESG) investing. ESG investors typically focus on nonfinancial indicators of corporate performance in the industrial economy, so they pursue information about a manufacturer's toxic waste disposal or the hiring practices of a law firm. Still, in significant ways the challenges confronted by ESG investors run parallel to those faced by AI ethics evaluators. In both cases, material nonfinancial information is sought. There is also a shared reliance on unstructured data: whether the subject is environmental damage or privacy harms, information must be extracted from narrative sources as opposed to spreadsheet numbers. Finally, ESG investors face the same cost and time pressures that currently apply to AI ethics audits.

Innovators in ESG finance have begun responding to the pressures with algorithms. Databricks, ESG Analytics, and Truevalue Labs (recently acquired by FactSet) exemplify the trend. In the case of Databricks, a two-pronged approach is applied to a company during evaluation. On one side, machine learning and natural language processing discerns a sense of the company's ESG claims and, on the other, a sentiment analysis of public information is employed to test the claims.[18] The results of this work will not be flawless, but they will be objective, rapid, and relatively inexpensive.

4. How can the ESG model be transferred to AI-intensive companies?

By applying natural language processing and machine learning in ways that parallel the ESG model, the two primary obstacles confronting AI ethics audits – cost and time – can be addressed. To execute that strategy, a collaboration between Pace University in New York City and the University of Trento in Italy has been initiated. Called *AI Applies AI Ethics to AI*, the project is currently in a

pilot stage, and initial work is focused on testing one company and its performance in a single category of algorithmic auditing. Because of its economic prominence, Meta's Facebook was selected as the company, and because of its humanistic significance in digital reality, privacy was selected as the category.

The envisioned process encompasses two steps. First, a model is constructed to read and understand Facebook's own claims and policies about user privacy. Then, a second model gathers and interprets public information that reflects Facebook's performance as measured against its own claims. So far, the first step has been completed.

Work began by annotating Facebook's legal documents and declarations. The texts included: Terms of Service, Data Policy, Cookies Policy, Open Source Privacy Policy, General Information on Protecting Privacy and Security. Taken from the Facebook website, these documents implicitly contain Facebook's privacy attitude. After tokenizing the text on the sentence level, the subsequent annotation labelled each sentence with one of the following: Commit to privacy, Violate privacy, Related to privacy, Declare opinion about privacy, Not applicable.

The first two categories emerged as the most germane to the work. Commitments to privacy generally entailed declarations limiting what Facebook would do with user information, for example the claim that personal data would not be sold to third parties. Violations of privacy were generally statements about how Facebook would use personal information, for example the declaration that advertising would be targeted based on personal characteristics discerned by user activity. Some sentences were related to privacy but seemed neutral with respect to Facebook, for example assertions that third parties had their own, distinct privacy policies. Sentences declaring an opinion about privacy were sparse assertions about the general importance of protecting user privacy. Finally, most sentences in the documents were non-applicable as they described company activities in areas outside of user privacy.

Completion of the annotation led to work in natural language processing. Analogously to the human annotator, an algorithm was designed to classify these sentences, automating the process.

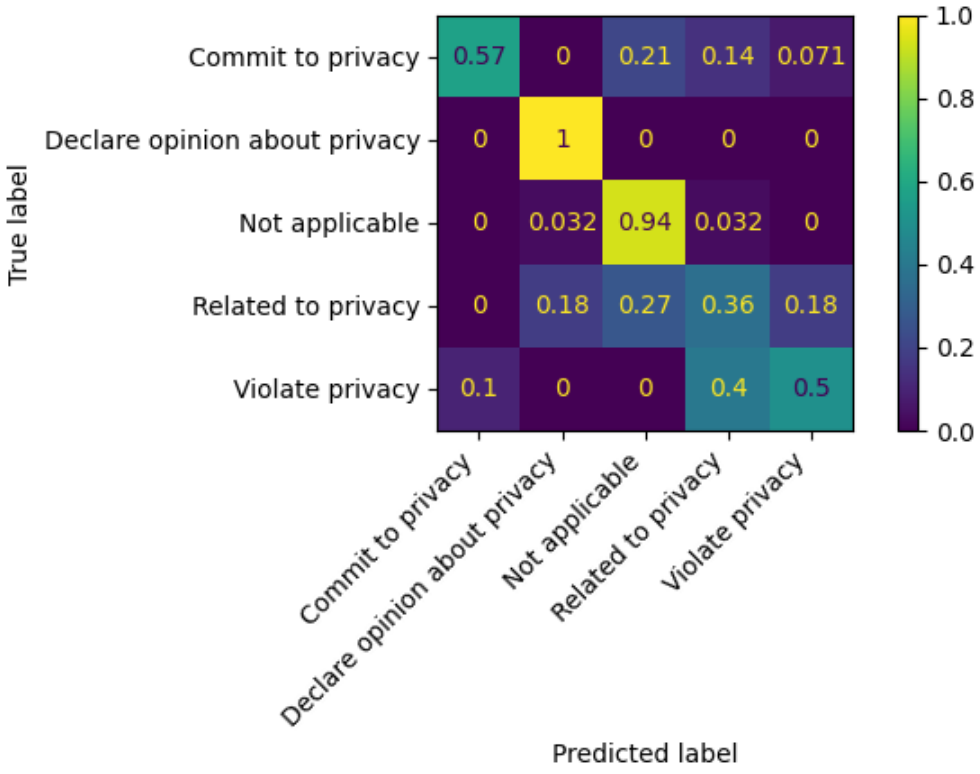
The initial step was to separate the dataset into three partitions: the training set, the development set, and the test set. The training set held 80% of the original information, and both the development and test sets were assigned half of the remaining data. The use of stratified distribution ensured that each class was equally represented in the partitions, regardless of their frequency disparities.

To form a baseline, a simple classifier that did not use any machine learning was developed. It predicted a class for each sentence of the input data and functioned by reference to human-selected keywords, with the keywords judged as most likely to suggest a sentence's belonging to one of the five classes (Commit to privacy, Violate privacy, Related to privacy, Declare opinion about privacy, Not applicable).

The next step was to develop a classifier that used machine learning techniques to execute the classification with superior performance. Machine learning methods use the given input data to increasingly enhance their performance at each training session. For that reason, they adapt well to new contexts. They were fundamental to this project because they allowed the interchangeability of the code, meaning that an algorithm that was developed using Meta's Facebook data as input can be later trained with data from other companies and still function well.

To implement the technique using our annotated data, several built-in machine learning classifiers of the python programming language were tested. The machine learning classifier that outperformed others in our tests was a Linear Support Vector Classifier (SVC). Its performance can be illustrated by a Confusion Matrix metric which displays the proportion of correct predictions of the algorithm according to the quantities and kinds of its mistakes. On the vertical axis lay the correct labels. On the horizontal axis lay the predicted labels. As a result, on the main diagonal, the percentage of correct predictions is shown for each class. In each line, some other positions are different than zero. Each of these positions shows the proportion of mistakes made by the machine by predicting the sentence of the true category, represented in the vertical axis, as a sentence from the category represented in the horizontal axis. Figure 1 illustrates the Confusion Matrix resulting from one of the configurations tested for the SVC.

Figure 1 - Confusion Matrix obtained by the Linear Support Vector Classifier (SVC), with regularization parameter C=1, and multiclass support using a One vs One scheme, when executed on the test set.



Results are imperfect but promising. Considering the best obtained results – which are also measured proportionally according to the importance of each class – the classifier suggests with more than 60% accuracy how sentences in a document can be divided in terms of their relevance for measuring privacy performance. Data from different sources and algorithmic refinement should increase accuracy and confidence that natural language processing and machine learning are effective in this scenario. Specifically, it should increase confidence that they are sensitive in their ability to locate sentences that correlate with significant privacy claims, and then divide that group between those sentences that relate how a company promises to respect user privacy, and those claims that relate how the company will exploit users' personal information. The algorithms should also be highly specific in the sense that they classify and set aside sentences that are only generally related to privacy, that only state an opinion about privacy, or that are unrelated to privacy.

The accomplishment of the classification step opens the way to the second and complementary part of the process, which is to devise a model that gathers information from public, unstructured narrative data and subsequently generates a measure of the company's adherence to its own privacy policies.

5. Conclusion: The future of AI ethics audits

The road to fully automatic AI ethics audits is long, but not endless. The *AI Applies AI Ethics to AI* project maps a way forward, and begins making progress toward an automated reality where a company's performance in terms of protecting user privacy is efficiently and constantly measured. From there, additional categories of algorithmic assessments including autonomy, fairness, and accountability can be included in the processing.

More broadly, the project is valuable for two reasons. First, regulatory requirements threaten to swamp technological innovation with algorithmic impact assessments that become expensive and lengthy when executed by human evaluators. Second, investors seek information to help identify technologies that work well on the human level and so stand a better chance of success in the marketplace. Natural language processing and machine learning potentially responds to both regulators and investors by producing knowledge about AI ethics inexpensively and quickly.

---

[1] Mökander, J., Axente, M., Casolari, F. *et al.* (2021). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds & Machines, available at* https://doi.org/10.1007/s11023-021-09577-4

[2] European Commission, Directorate-General for Communications Networks, Content and Technology. (2019). Ethics guidelines for trustworthy AI, available at https://data.europa.eu/doi/10.2759/177365

[3] Press release. (2022). Wyden, available at https://www.wyden.senate.gov/news/press-releases/wyden-booker-and-clarke-introduce-algorithmic-accountability-act-of-2022-to-require-new-transparency-and-accountability-for-automated-decision-systems

[4] Kaye, Kate. (2022). This Senate bill would force companies to audit AI used for housing and loans. Protocol. February 8, 2022. Available at https://www.protocol.com/enterprise/revised-algorithmic-accountability-bill-ai

[5] Mishra, Subodh. (2020). ESG Matters. Harvard Law School Forum on Corporate Governance. Tuesday, January 14, 2020, available at https://corpgov.law.harvard.edu/2020/01/14/esg-matters/

[6] Cornell, Bradford. (2021). ESG preferences, risk and return. *European Financial Management* 27.1 (2021): 12-19.

[7] Chant, Tim De. (2022). Meta may be forced to shutter Facebook, Instagram in EU Stringent data privacy rules under GDPR have the company in a tight spot. Ars Technica. 2/7/2022, available at https://arstechnica.com/tech-policy/2022/02/meta-may-be-forced-to-shutter-facebook-instagram-in-eu/

[8] C. Laurell and C. Sandström. (2022). "Social Media Analytics as an Enabler for External Search and Open Foresight—The Case of Tesla's Autopilot and Regulatory Scrutiny of Autonomous Driving," in *IEEE Transactions on Engineering Management*, vol. 69, no. 2, pp. 564-571, April 2022, doi: 10.1109/TEM.2021.3072677.

[9] European Commission, Directorate-General for Communications Networks, Content and Technology. (2019). Ethics guidelines for trustworthy AI, available at https://data.europa.eu/doi/10.2759/177365

[10] Data Ethics Commission of the Federal Government. (2019). Opinion of the German Data Ethics Commission. Federal Ministry of the Interior, Building and Community of Germany and Federal Ministry of Justice and Consumer Protection of Germany.

[11] Mökander, J., Floridi, L. (2021). Ethics-Based Auditing to Develop Trustworthy AI. *Minds & Machines* 31, 323–327 (2021). https://doi.org/10.1007/s11023-021-09557-8

[12] Brusseau, J. (2020). What a Philosopher Learned at an AI Ethics Evaluation. AI Ethics Journal, 2020, 1(1)-4, https://doi.org/10.47289/AIEJ20201214

[13] R. V. Zicari et al. (2021). A Process to Assess Trustworthy AI, in IEEE Transactions on Technology and Society, vol. 2, no. 2, pp. 83-97, June 2021, doi: 10.1109/TTS.2021.3066209.

[14] Brusseau, J. (2020). What a Philosopher Learned at an AI Ethics Evaluation. AI Ethics Journal, 2020, 1(1)-4, https://doi.org/10.47289/AIEJ20201214

[15] Glauner, Patrick. (2022). An Assessment of the AI Regulation Proposed by the European Commission. In book: The Future Circle of Healthcare: AI, 3D Printing, Longevity, Ethics, and Uncertainty Mitigation Publisher: Springer.

[16] Brusseau, J. (2022). From the ground truth up: doing AI ethics from practice to principles. *AI & Soc* (2022). https://doi.org/10.1007/s00146-021-01336-4

[17] AI Ethics Decentralized and Accelerated. (2021). United Nations Educational, Scientific and Cultural Organization (UNESCO) Information For All conference. 28 September 2021.

[18] Amend, Antoine. (2020). A Data-driven Approach to Environmental, Social and Governance July 10, 2020 in Databricks Engineering Blog, available at https://databricks.com/blog/2020/07/10/a-data-driven-approach-to-environmental-social-and-governance.html