DOI:10.1145/3625252

N. Diakopoulos, C. Trattner, D. Jannach, I. Costera Meijer, and E. Motta

Computing Ethics
Leveraging Professional Ethics for Responsible AI

Applying AI techniques to journalism.

ARTIFICIAL INTELLIGENCE (AI) is proliferating throughout society, but so too are calls for practicing *Responsible AI.*[4] The ACM Code of Ethics and Professional Conduct states computing professionals should contribute to society and human well-being (General Ethical Principle 1.1), but it can be difficult for a computer scientist to judge the impacts of a particular application in all fields. AI is influencing a range of social domains from law and medicine to journalism, government, and education. Technologists do not just need to make the technology work and scale it up, they must make it work while also being responsible for a host of societal, ethical, legal, and other human-centered concerns in these domains.[11]

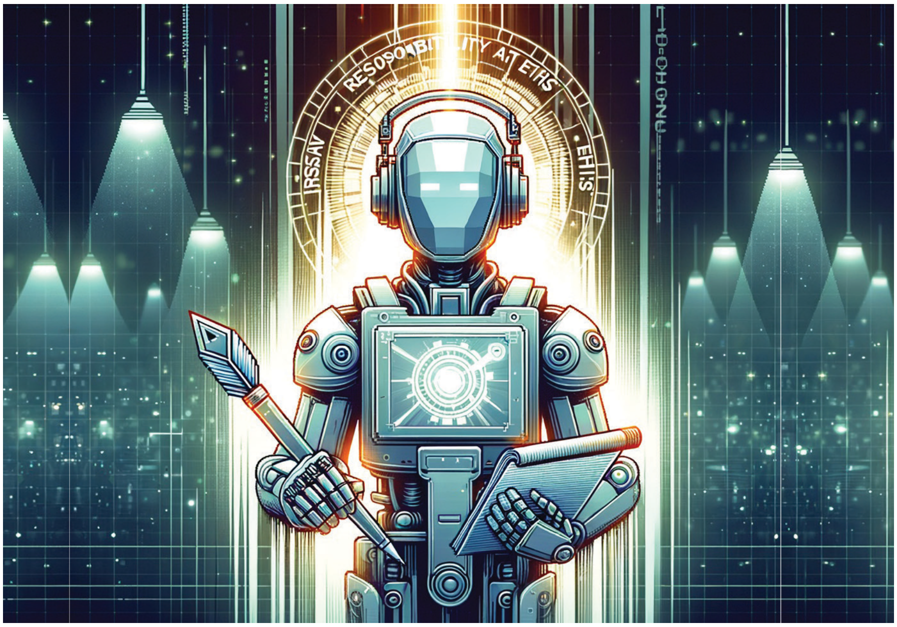


IMAGE CREATED BY ANDRIJ BORYS ASSOCIATES WITH THE ASSISTANCE OF SHUTTERSTOCK.AI

There is no shortcut to becoming an expert social scientist, ethicist, or legal scholar. And there is no shortcut for collaborating with those experts and doing careful evaluations with stakeholders to understand the impacts of technology. But there is a way to at least jumpstart your knowledge and enrich your understanding of what it takes to responsibly implement technology in a social domain: domain-specific professional codes of conduct.

Professional codes of conduct are shortcuts for understanding commonly held values in a domain. They articulate general expectations for responsible practice and facilitate the understanding of issues arising around that practice. Similar to the computing field, medicine, law, architecture, journalism, and plenty of other professions have spent years working through and crystallizing what it means to act ethically in their particular domain of society.[5]

Technologists can take these codes as a starting point for informing the design, engineering, and evaluation of responsible AI systems that comply with the ACM Code of Ethics and Professional Conduct. Such codes are not checklists that will automatically make the AI (or the technologist creating that AI) "responsible." Rather, the reasoning and values embodied in such codes can help guide technologists toward more responsible choices in their practice as they develop AI-based systems for these domains.

We illustrate this idea in the context of applying AI techniques in the domain of journalism, which is familiar to most readers but differs sufficiently from academic publication in that more guidance is required than is provided by the ACM Code of Ethics. We first introduce some key professional values in journalism, elaborating a design-build-evaluate process for incorporating those values into a system, and then show how this process applies specifically to an algorithmic content-curation feed.

From Journalism Ethics to Responsible AI for Media

Professional journalism ethics has long grappled with "the responsible use of the freedom to publish."[12] Although journalism ethics is ever evolving to address new social and technical conditions, journalistic codes of conduct embed institutionalized ethical principles of proper behavior, which can guide computing professionals designing and developing new media technologies.[8] The Society of Professional Journalists (SPJ)[2] offers four main tenets: seek truth and report it, minimize harm, act independently, and be accountable and transparent (see the table).

These affirmative duties, reflecting deeply held domain values, can inform the design of an ethical media system. The accompanying figure shows how the domain-specific values in journalism can inform the design, build, and evaluation phases of engineering, show how principles and norms of responsible practitioners can be reflected in the data, algorithms, metrics, and organizational processes built into new responsible AI systems.

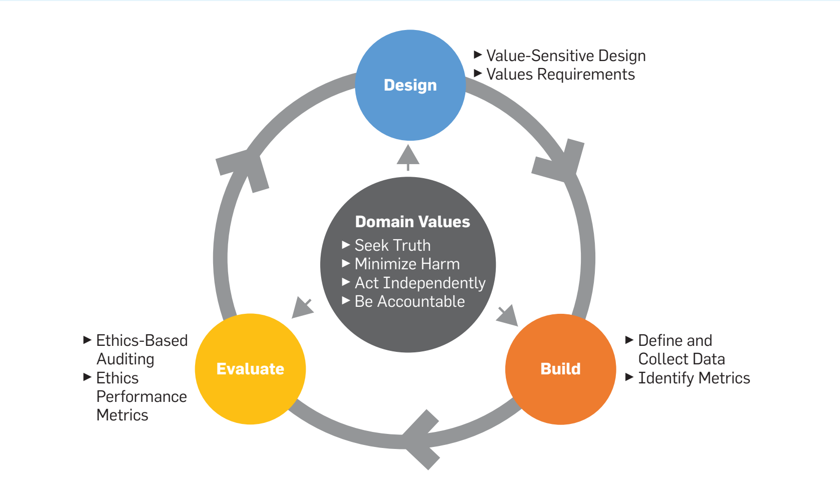
By first identifying key domain values, technologists can incorporate them as requirements during the design process, generating ideas for technical and organizational features that support those values.[7] Value-sensitive design methods, which "account for human values in a principled and systematic manner throughout the technical design process," can be used to help translate domain values into design features.[6]

Technologists must build their systems to reflect the required values identified during design. Here, we focus on AI systems using machine learning where defining and collecting the right datasets and finding the right metrics for training and evaluation are crucial. For instance, the editorial algorithm that curates Swedish Radio's audio feeds is trained by experienced news editors rating content on whether it meets the organization's public service goals. Aligning data and metrics to values is a non-trivial challenge. Principles can be vague and multivalent, hiding potential ethical conflicts (for example, around notations of "fairness").[9] Success at the build stage requires clear definitions of values so that training data can be appropriately operationalized.

The four main principles of journalism ethics as described by the Society of Professional Journalists (SPJ).

Principle	Description
Seek Truth	Ensure the accuracy of published information, including providing interpretive context to avoid distortion, attributing to sources, and labeling commentary.
Minimize Harm	Interact with stakeholders (sources, subjects, the public, and impacted communities) and balance the consequences of seeking and publishing information against the harm that may cause.
Act Independently	Manage and/or disclose conflicts of interest so they can reflect the public interest without any real or perceived external influence, favoritism, or self-interest.
Be Accountable	Take responsibility for what is published and explain how you know what you know and why you are publishing something. Promptly and prominently acknowledge, respond to, and correct issues.

Responsible design and engineering of AI for media incorporating journalism-specific domain values that inform an iterative design, build, and evaluation process.



Finally, technical developers must evaluate the value alignment of the overall system. To do this, they should implement ethics-based auditing (EBA), which is a method to assess a system's "consistency with relevant principles or norms."[10] EBA supports responsible AI development and continuous improvement by evaluating a system's impact on relevant design values. There may be ethical performance metrics that can demonstrate that a value is being upheld. For instance, platforms use machine learning to detect child sexual abuse material and then report the volume of content "actioned" (for example, removed).[1] Tracking system performance against ethical performance metrics is useful internally for improving the system, and can also inform the public as part of transparency disclosures that support accountability and build trust in the system.

An Application to Algorithmic Content Curation

A key challenge for responsible AI is to translate abstract ethical principles into real systems. As an example, we illustrate how the SPJ principles can inform the engineering of an ethical algorithmic newsfeed for a social media platform. These values can inform the design, build, and evaluation process to ensure the principles are enacted in the sociotechnical embodiment of a feed.

To seek truth at scale in a newsfeed, designers and developers must ensure the accuracy of information in the feed. Sociotechnical design features include algorithms to filter out disinformation, amplification supervisors to manually review content receiving exceptional levels of attention, and content labels for sources and opinions. To build such features, data will often need to be defined and collected to train models. For instance, fact-checking specialists might annotate content for training and improving models to reduce misinformation. For model evaluation, feed operators should track feed-quality metrics by sampling and evaluating quality both automatically and manually using a systematized rubric.

Content moderation is required to help minimaze harm that content can cause to individuals. Data scientists might design and develop classifiers that automatically detect content about non-public victims of crime, abuse, bullying, or violations of privacy and block its amplification. Similar classifiers could filter out content for vulnerable people to avoid, for example, content reinforcing depression, eating disorders, or self-harm, and then evaluate the impact of that filtering. Harms to society, such as affective polarization, could be evaluated with key metrics (for example, engagement across political lines) to adjust the system over longer timeframes.

Independence in the feed demands conflicts of interest are managed and disclosed so the public interest is prioritized. Platforms could develop a database of sources, how they are funded (for example, commercial, non-profit, sponsored, and so forth), who owns them (for example, corporations, hedge funds, foreign entities), whether they are paid by the feed operator, and so on. To keep the database updated, new sources could be automatically identified based on feed exposure, with details filled in by trained curators. Source information could be disclosed in the user interface (for example, with labels) to clarify where information comes from and the relationship between the feed operator and the source. Evaluations of the newsfeed could assess exposure to different types of sources, such as foreign media.

Upholding the principle of accountability requires the feed algorithm provide transparency including things like datasheets and thick descriptions of system processes.[3] Periodic disclosures of internal evaluations (for example, ethics-based audits) could clarify how the feed operator is upholding the values of seeking truth, minimizing harm, and maintaining independence with key performance indicators that can identify lapses (for example, increased exposure to disinformation). Ultimately the company is accountable for how the system functions and creates a user experience, even if they do not fully control all the components that contribute to that experience. "Algorithm explainer" roles could be created with full access to system information and responsibility to explain system behavior in case of errors or malfunctions. These technical workers would release public reports or respond directly to individuals seeking redress.

Technical experts have an essential role to play in designing and engineering AI systems to be socially responsible.

Conclusion

Technical experts have an essential role to play in designing and engineering AI systems to be socially responsible. While technologists cannot all be experts in social science, ethics, or law, they can effectively leverage institutionalized domain values as generative design tools for feature suggestions, and as guides for developing and evaluating the value-aligned implementation of a system. Domain values should be seen as a way to coarsely aim the process in the right direction, but it is important to emphasize that technologists will also need to iterate on systems through deep human-centered work with various stakeholders in the domain. And as jurisdictions around the world move to regulate AI, such as in the DSA and AI Acts in the E.U., the same techniques for aligning with domain values will facilitate developing systems that are adherent to broader social values, such as fundamental rights.

By centering the values of the domain present in professional codes of conduct we can leverage the received wisdom of thousands of professionals who have already worked through some of the stickiest of ethical issues and ambiguities of responsible practice. Just as we have demonstrated here for journalism, other domains (for example, law, medicine) would similarly benefit by leveraging domain-specific ethics codes in designing and aligning responsible AI systems, including for fine-tuning some of the latest generative AI models, such as those powering ChatGPT, to align them with expectations of responsible behavior in particular domains. Let's be sure that as designers and engineers we incorporate such wisdom, and build our technologies to aspire to and embody domain values of responsible practice and of The ACM Code of Ethics and Professional Conduct.

References

1. *Automated Content Moderation: A Primer*. Stanford Cyber Policy Center, (2022); https://bit.ly/3LT2qyr.
2. Brown, F. *Media Ethics: A Guide for Professional Conduct. Society of Professional Journalists Foundation*. SPJ Code of Ethics (2020); https://www.spj.org/ethicscode.asp.
3. Diakopoulos, N. and Koliska, M. Algorithmic transparency in the news media. *Digital Journalism 5*, (2016), 7.
4. Dignum, V. Responsibility and artificial intelligence. *Oxford Handbook of Ethics and AI*. M. Dubber and F. Pasquale, (Eds). Sunit Das. 2020.
5. Frankel, M.S. Professional codes: Why, how, and with what impact? *J. Business Ethics 8*, 2–3 (1989); 10.1007/bf00382575
6. Friedman, B. and Hendry, D.G. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, 2019.
7. IEEE Standard Model Process for Addressing Ethical Concerns during System Design. IEEE Std 7000-2021. (2021), 1–82; 10.1109/ieeestd.2021.9536679
8. McBride, K. and Rosenstiel, T. *The New Ethics of Journalism: Principles for the 21st Century*. CQ Press. 2014; *Ethics for Digital Journalists: Emerging Best Practices*. L. Zion and D. Craig, (eds). Routledge, 2015.
9. Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence 1*, 11 (2019), 501–507.
10. Mökander, J. et al. Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics 27, 4* (2021); 10.1007/s11948-021-00319-4
11. Trattner, C. et al. Responsible media technology and AI: Challenges and research directions. *AI and Ethics*. (2021).
12. Ward, S. *Disrupting Journalism Ethics (Disruptions)*. Taylor and Francis, 2019.

Nicholas Diakopoulos (nad@northwestern.edu) is a professor of communication studies and computer science, Northwestern University, Evanston, IL, USA.

Christoph Trattner (christoph.trattner@uib.no) is a professor of information science and media studies, University of Bergen, Bergen, Norway.

Dietmar Jannach (Dietmar.Jannach@aau.at) is a professor of computer science, University of Klagenfurt, Wörthersee. Austria.

Irene Costera Meijer (icostera.meijer@vu.nl) is a professor of journalism studies, Vrije University Amsterdam, Netherlands.

Enrico Motta (Enrico.Motta@open.ac.uk) is a professor of knowledge technologies, Open University, London, U.K.

This work is funded in part by MediaFutures partners and the Research Council of Norway (grant number 309339).