AI Ethics Canvas: Guidebook

The AI Ethics Canvas is a tool for helping you and your team think through ethical issues that might arise in the development and deployment of an artificial intelligence–enabled system or machine learning model. It combines various different perspectives from philosophical ethics, professional ethics in computing, and emerging AI ethics frameworks.

For a list of references and further reading, see Appendix D – Bibliography.

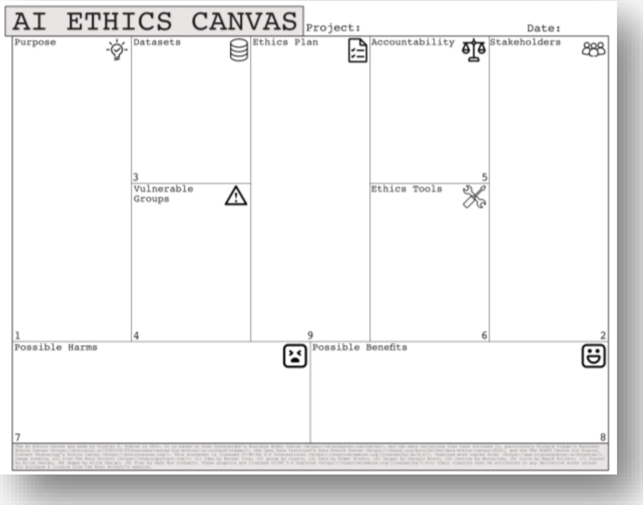
The AI Ethics Canvas

The AI Ethics Canvas is divided into nine cells to be filled in. It should be completed as part of the initial steps in your design cycle – corresponding to the Discover phase, or the initial divergent thinking phase, in the Double Diamond framework. Revisit the AI Ethics Canvas at each subsequent step of the design cycle – in the Double Diamond framework, Define, Develop, Deliver, and Evaluate.

Before beginning, familiarize yourself with the layout of the canvas and consider what additional resources you may need to assemble in order to complete it. For example, you may want to have a few browser tabs open with your company's mission statement; your databases; any relevant policies, regulations, or contracts; and so on.

The nine sections are as follows. Each is detailed in subsequent sections of this guidebook:

1. Purpose
2. Stakeholders
3. Datasets
4. Vulnerable Groups
5. Accountability
6. Ethics Tools
7. Possible Harms
8. Possible Benefits
9. Ethics Plan



The AI Ethics Canvas can be completed by an individual, but it is more effective to complete it in a group. Diverse perspectives are essential to catching bias and other ethical issues. Consider having a meeting of at least an hour with relevant team members. If possible, include external stakeholders, members of other relevant departments, and executives.

Before beginning to fill in the first cell, write the name of the project and the date. Decide as a group whether to include the names of all those present at the meeting, or whether to use the Chatham House Rule (i.e., anyone who attends the meeting is allowed to use any information provided or created in the discussion, but is not allowed to attribute any comment to any particular attendee) or a similar procedure to ensure mutual trust and safety for all participants.

Each cell has several questions to consider when filling it in. You don’t necessarily have to answer them all – and not all of them may be applicable to your situation. But you should consider each of the questions provided, and how your answers to one set of questions shapes your answers to the others.

You’ll also find that sometimes after answering one set of questions, new considerations occur to you that are relevant to earlier questions. Feel free to jump back and forth through the cells as new things come to light! Like design, ethics planning is an iterative process.

Finally, don’t feel that these are the *only* important questions. If you think of other considerations, include them! Make the AI Ethics Canvas your own – and let the creator know your experiences and suggestions by emailing `contact@trystangoetze.ca`.

1. Purpose

Consider the following questions:

- What kind of AI-enabled system are we proposing to create?
- For what *internal* purpose are we creating this AI system? What problem(s) will this system solve *for us*?
- For what *external* purpose are we creating this AI system? What problem(s) will this system solve for our customers, clients, stakeholders, etc.?

Created by Berkah Icon
from Noun Project

2. Stakeholders

Consider the following questions:

- Who are your primary stakeholders, in general? Consider your customers or clients, and anyone else who might be affected by your business decisions.
- Of your stakeholders, who will be interacting directly with the AI system? Alternatively, who will be directly impacted by the system's behaviour?
- Which stakeholders have you already communicated with about the project? Which stakeholders have you not consulted yet?
- Are there any stakeholders you hadn't yet considered when thinking about the purposes this AI system might serve for them? If yes, go back to (1) before continuing.

Created by vigom
from Noun Project

3. Datasets

Consider the following questions:

- What data do we need to train and test the AI system we have proposed?
- What data do we already have at our disposal?
- What data do we need to obtain?
- Are there privacy or consent issues with using these data?
- Are the data representative? Are they biased?

Created by Kimmi Studio
from Noun Project

4. Vulnerable Groups

Consider the following questions:

- Of our stakeholder groups, are any historically or currently marginalized with regard to race, ethnic origin, sex, gender, sexuality, disability, age, religion, language, indigenous status, or other equity-deserving categories?
- Due to their marginalization, are any of these groups particularly vulnerable to errors, bias, or harm due to the AI system we propose to develop?
- Have people from those groups been consulted as part of the development of the AI system? If yes, what was their

Created by Gonzalo Bravo
from Noun Project

perspective? If not, pause and make plans to hold such consultations before continuing.

5. Accountability

Consider the following questions:

Created by Mutualism from Noun Project

- How will we ensure that any decisions taken by or with the aid of the AI system will be explainable to our stakeholders?
- Is our use of AI and datasets compliant with relevant privacy and data control legislation (e.g. GDPR, PIPEDA)?
- If a stakeholder or regulator asks a question about how their data are used, are we prepared to offer an explanation of which data are collected or processed, for what purpose, and how long they will be retained? If not, pause and make plans for how to handle such a situation.
- Who is responsible for data security in connection with this project?

6. Ethics Tools

Consider the following questions:

Created by Maxim Kulikov from Noun Project

- Will we be using ethics tools for ensuring fairness and lack of bias in our data and in our models' outputs (e.g., the Aequitas Bias and Fairness Audit)?
- Are our datasets and machine learning models accompanied by explanatory information slips (e.g., datasheets for datasets, model cards for model reporting)?
- Will we use an assessment tool to ensure that our AI systems meet the standards of AI ethics principles (e.g. the Responsible AI Design Assistant)?
- Are there any other ethics tools that would be useful to this project?

See Appendix D – Bibliography for links to these and other tools.

7. Possible Harms

Consider the following questions:

- What forms of physical harm could arise due to our AI system?
- What forms of emotional or psychological harm could arise due to our AI system?
- Could our AI system unfairly deny people access to economic opportunities?
- Are there circumstances where our AI system could infringe on people’s human rights (e.g., right to dignity, liberty, privacy)?
- What is the environmental impact of developing and using the AI system?
- Could our AI system manipulate people or contribute to social harms (e.g., misinformation, propaganda, inequity)?
- Will implementing this AI system replace any human workers?
- Rate each of these harms based on how significant they are, and how likely they are to occur.

Created by Alice Design
from Noun Project

Before moving on, ask if there are any perspectives, particularly those of historically marginalized groups, that haven’t yet been considered in this harms assessment.

8. Possible Benefits

Consider the following questions:

- What economic or efficiency benefits might this AI system provide for our organization? For our stakeholders?
- What emotional or psychological benefits might this AI system provide for our employees? For our stakeholders?
- Will our AI system promote greater equity for historically marginalized groups?
- Are there other benefits that may be produced by our AI system (e.g., health, environmental, educational)?

Created by Alice Design
from Noun Project

Compare your answers in this cell to your answers under Possible Harms. Does it look like the AI system you’re developing will have a net positive, net negative, or neutral impact on society?

9. Ethics Plan

Synthesize the contents of the other cells to create a list of concrete next steps to ensure that the current and following phases of your design and development cycle is ethically responsible.

Created by Hafiz Nur Lutfianto from Noun Project

Assign specific people or teams to particular AI ethics tasks, such as data security, privacy liaison, stakeholder meeting facilitation, completion of AI ethics tools.

Put timescales on the different tasks you identified above. For example, When will you meet with stakeholders? When will the ethics tools be completed? Ensure that these tasks are explicitly part of your project pipeline.

Plan to revisit this canvas periodically, as major milestones are reached in your project's schedule. Consider any new information that has come to light during the design and consultation process and revise accordingly.

Finally, plan ahead to review and evaluate the AI system's performance some time after implementation, taking into account not just where it is achieving its internal and external purposes, but also whether any unanticipated ethical issues have arisen.

Appendices

The following sections contain bibliographic and legal information. You should familiarize yourself with Appendix C – Licence before using the AI Ethics Canvas.

A. Credits

This is version 0.2 of the AI Ethics Canvas. Find the most up-to-date version here: `https://www.trystangoetze.ca/AIcanvas`

The AI Ethics Canvas was created in 2021 by Trystan S. Goetze, Ph.D., in consultation with Katrina Ingram and Ethically Aligned AI, Inc. It is based on Alex Osterwalder's Business Model Canvas and the many variations that have followed it. In particular, the following were important influences on this canvas and guidebook:

- Business Model Canvas, Alex Osterwalder, `https://strategyzer.com/canvas`

- Business Ethics Canvas, Richard Vidgen, `https://ethicalai.ai/2020/05/09/business-canvas-for-ethical-ai-richard-vidgen/`
- Data Ethics Canvas, Open Data Institute, `https://theodi.org/article/the-data-ethics-canvas-2021`
- Ethics Canvas, ADAPT Centre for Digital Content Technology, `https://ethicscanvas.org/`

Learn more about Trystan's work here: `https://www.trystangoetze.ca/`

Learn more about Ethically Aligned AI here: `https://www.ethicallyalignedai.com/`

B. Artwork

The graphics used on the AI Ethics Canvas and in this guidebook are from the Noun Project, a collection of community-created vector graphics, `https://thenounproject.com`. Each graphic is licensed CC-BY 3.0 `https://creativecommons.org/licenses/by/3.0/`. Redistribution of these graphics without inclusion of these attributions is prohibited, unless you purchase a licence from the creators through the Noun Project's website.

1. Idea by Berkah Icon, `https://thenounproject.com/icon/1291989/`
2. group by vigorn, `https://thenounproject.com/icon/2350086/`
3. Data by Kimmi Studio, `https://thenounproject.com/icon/1832609/`
4. Danger by Gonzalo Bravo, `https://thenounproject.com/icon/50917/`
5. Justice by Mutualism, `https://thenounproject.com/icon/3778371/`
6. tools by Maxim Kulikov, `https://thenounproject.com/maxim221/collection/hand-tools/?i=943563`
7. Scared by Alice Design, `https://thenounproject.com/icon/2546414/`
8. Happy by Alice Design, `https://thenounproject.com/icon/2562064/`

C. Licence

The AI Ethics Canvas and this Guidebook are provided under a CC-BY-SA 4.0 International licence `https://creativecommons.org/licenses/by-sa/4.0/`. In brief, you are welcome to make and distribute copies of these works for any purpose, so long as any copies and derivative works are distributed under the same licence or a substantively similar licence (such as the GNU General Public Licence). Additional copies of the AI Ethics Canvas and this guidebook are available from `https://www.trystangoetze.ca/AIcanvas/`. If you use this canvas, please let the creator know about your experience by sending an email to: `contact@trystangoetze.ca`.

D. Bibliography

The following articles and books were consulted in the making of this Guidebook. Consider reading them yourself!

Abrassart, Christophe, Yoshua Bengio, Guillaume Chicoisne, Nathalie de Marcellis-Warin, Marc-Antoine Dilhac, Sébastien Gambs, Vincent Gautrais, Martin Gibert, Lyse Langlois, François Laviolette, Pascale Lehoux, Jocelyn Maclure, Marie Martel, Joëlle Pineau, Peter Railton, Catherine Régis, Christine Tappolet, and Nathalie Voarino. 2018. *The Montréal Declaration for a Responsible Development of Artificial Intelligence*. `https://www.montrealdeclaration-responsibleai.com/the-declaration`

AIGlobal. 2020. Responsible AI Design Assistant. `https://oproma.github.io/rai-trustindex/`

Algorithm Watch. 2020. AI Ethics Guidelines Global Inventory. `https://inventory.algorithmwatch.org/`

Association for Computing Machinery. “ACM Code of Ethics and Professional Conduct.” `https://ethics.acm.org/code-of-ethics/`

ADAPT Centre for Digital Content Technology. 2017. “Online Ethics Canvas.” `https://ethicscanvas.org/`

Bender, Emily M., and Batya Friedman. 2018. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.” *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. `https://doi.org/10.1162/tacl_a_00041`

Broad, Ellen, Amanda Smith, and Peter Wells. 2017. *Helping Organizations Navigate Ethical Concerns in the Data Practices*. The Open Data Institute, London, UK. `https://www.scribd.com/document/358778144/ODI-Ethical-Data-Handling-2017-09-13`

Calvo, Rafael, and Dorian Peters. n.d. Responsible Tech Design: Tools and methods for more ethical practice in technology design. `https://www.responsibletechdesign.com/`

Cassidy, Doug, Alex Buck, Dhanashri Kshirsagar, and Harmony Mabrey. 2020. Harms Modeling. `https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/`

Kasia S. Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2020. “The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence.” In *NeurIPS 2020 Workshop on Dataset Curation and Security*. Neural Information Processing Systems Foundation, San Diego, CA, 7 pages.

The Design Council. 2015. What is the framework for innovation? Design Council’s evolved Double Diamond. `https://www.designcouncil.org.uk/news-opinion/what-framework-innovation-design-councils-evolved-double-diamond`

Floridi, Luciano. 2019. “Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical.” *Philosophy & Technology* 32: 185–93.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christophe Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.” *Minds and Machines* 28: 689–707. `https://doi.org/10.1007/s11023-018-9482-5`

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. “Datasheets for Datasets.” In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning* (Stockholm,Sweden). ACM, New York, NY, 24 pages. `https://arxiv.org/pdf/1803.09010.pdf`

Gilligan, Carol, 1982, *In a Different Voice: Psychological Theory and Women's Development*, Cambridge, MA: Harvard University Press.

Kant, Immanuel. 1993/1785. *Grounding for the Metaphysics of Morals*. Hackett, Indi anapolis, IN.

Mill, John Stuart. 1879. *Utilitarianism*. Longmans, Green, and Co., London, UK.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. “Model

Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) (FAT* '19). Association for Computing Machinery, New York, NY, 220–229. `https://doi.org/10.1145/3287560.3287596`

Noddings, Nel. 2013. *Caring: A Relational Approach to Ethics and Moral Education*, 2nd ed. Berkeley and Los Angeles: University of California Press.

Norlock, Kathryn. 2019. "Feminist Ethics." *The Stanford Encyclopedia of Philosophy*, Summer 2019 ed. Edward N. Zalta (ed.). `https://plato.stanford.edu/archives/sum2019/entries/feminism-ethics/`.

Osterwalder, Alexander. 2020. "Business Canvas - Business Models and Value Propositions." `https://www.strategyzer.com/canvas`

Peters, Dorian, Karina Vold, Diana Robinson, and Rafael A. Calvo. 2020. "Responsible AI—Two Frameworks for Ethical Design Practice." *IEEE Transactions on Technology and Society* 1(1): 34–47.

Rawls, John. 1999. *A Theory of Justice*, revised ed. Cambridge, MA: Harvard University Press.

Responsible Artificial Intelligence Institute. n.d. Responsible AI Community Portal. `https://portal.ai-global.org/`

Richards, John, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. 2020. "A Methodology for Creating AI FactSheets." arXiv:2006.13796 [cs.HC] `https://arxiv.org/abs/2006.13796`

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2019. "Aequitas: A Bias and Fairness Audit Toolkit." arXiv:1811.05577 [cs.LG] `https://arxiv.org/abs/1811.05577`

Tavani, Herman T. 2015. *Ethics and Technology: Ethical Issues in an Age of Information and Communication Technology*. 5th ed. New York: John Wiley & Sons.

University Center for Human Values and Center for Information Technology Policy. 2021. Princeton Dialogues on AI and Ethics Case Studies. Princeton University. `https://aiethics.princeton.edu/case-studies/`

Vidgen, Richard, Giles Hindle, and Ian Randolph. 2020. "Exploring the ethical implications of business analytics with a business ethics canvas." *European Journal of*

Operational Research 281(3): 491–501. `https://doi.org/10.1016/j.ejor.2019.04.036`