Ethics and Trustworthy Foundation Models

Brian Patrick Green, Ph.D.
Director of Technology Ethics
Markkula Center for Applied Ethics
Santa Clara University
SMC-IT/SCC 15 July 2024

Abstract

What does it mean for something to be “trustworthy”?

At the very least, it must be both ***technically trustworthy*** - it does what it is supposed to do - and ***ethically trustworthy*** - it does not violate ethical ideals necessary for trust (such as violating rights, deceiving, harming, exploiting users, etc.).

This talk will explore linkages between AI and trust and present some ethical tools for thinking about and building trustworthy technology.

Markkula Center
for Applied Ethics
at Santa Clara University

Outline

I. Questions about Trust in Technology & AI: Delineations of the Solvable

II. Ethical Solutions

1. The Markkula Center Framework for Ethical Decision Making
2. Model Cards
3. Principles for Transparency
4. Ethics in Technology Practice

I. Ethics Is about Good Judgement

- Everyone should know how to make good decisions
- Tech empowers people to do new things. At the forward edges of human action people can act in ways that laws might not cover, but ethics does
- Ethics increases overall levels of trust in society by increasing *trustworthiness*

I. Technology and Trust

1) Technological products should be technically trustworthy:
   - They are tools that should do what they are supposed to do

2) Technological products should be ethically trustworthy:
   - They should have the user's best interests and the common good in mind, not exploit, deceive, violate, or otherwise harm people

The above are the minimum! Necessary, but not sufficient, for trust. Even if both are the case, technology can still create social distrust

I. With Tech, There Is a Third Source of Distrust…

- Simply adding 1) *functional* 2) *ethical* technology does not necessarily help to increase social trust

- As a side effect, it actually may harm social trust. Why?

Markkula Center
for Applied Ethics
at Santa Clara University

I. Why Does Tech Harm Social Trust?

More Technology	**=**	**More Power**
More Power	**=**	**More Choices**
More Choices	**=**	**More Responsibility**
More Responsibility	**=**	**More Need for Ethics**

- We were previously ***involuntarily*** constrained by our ***weakness***
- Now we must learn to be ***voluntarily*** constrained by our ***judgment***

- In other words, technological power turns socio-technical ***constants*** into ***variables*** (B. Srinivasan)

Markkula Center
for Applied Ethics
at Santa Clara University

I. Technological Power and Trust

When a constant becomes a variable it becomes a *choice* and we become *responsible* for it

- *Former constant*: no nuclear weapons, no nuclear winter, etc.
- *Former constant*: no space travel, no space debris, etc.
- *Former constant*: no anthropogenic climate change, no question of climate engineering, etc.
- *Former constant*: no "intelligent" products like AI, etc.

There are probably some constants that should not be turned into variables…

I. How Trust Is Harmed by Technology

- Constants can be trusted – *even if not that great* (death and taxes…), at least people know what to expect: *there is certainty*

- *Variables cannot be trusted* – even with great opportunities, the uncertainty and risk impede trust
  - Even if you trust the *tech product*, and trust the *person*, the *situation* may be untrustworthy, or even the thought of someone else's situation may inspire *worry* or "concern"
    - "I heard this happened to someone… will this happen to me?"

Markkula Center
for Applied Ethics
at Santa Clara University

I. How Trust Is Harmed by Technology

- As more choices become available, *uncertainty increases*, harming trust, *and when the right choices are not made* social trust is harmed again, a *double harm* to trust

- ***Variables cause WORRY****….* and people hate worrying. Worry indicates lack of trust

- Yet variables are also *opportunities* for those of more sanguine disposition

I. Technology and Trust

1) Technological products should to be technically trustworthy: they should do what they are supposed to do

2) Technological products should be ethically trustworthy: they should not exploit, deceive, violate, or otherwise harm people

3) But even if both technically and ethically trustworthy, *socially and psychologically*, technological products may still harm trust simply because they create uncertainty and worry

I. Tech and Trust and Science

- Social worry potentially affects everyone subjected to technological change and cannot be addressed by any individual user or producer
- Social worry can only be stopped by freezing the variable back into a constant by using ethical norms or law
- When technological power changes “impossible” problems into “hard” problems, it changes a constant into a variable. When society (whom exactly?) turns the tech back into a constant the “hard” problem is thenceforth “easy,” and accepted

Markkula Center
for Applied Ethics
at Santa Clara University

I. So… The Delineation of the Problem

Nobody here is going to solve the social-psychological problem of distrust due to technologically-induced worry related to constants becoming variables and not turning back into constants fast enough – at least not any time soon – though we can all *help* in this endeavor by laying the foundations: technically and ethically trustworthy systems

Technically trustworthy systems that function as expected? That is something people here can do

Ethically trustworthy systems that benefit society? That is something people here can do

II. Ethical Solutions

You all are the technical experts, not me, so I can do nothing there

But I can share ethical tools for creating ethically better AI systems

1. The Markkula Center Framework for Ethical Decision Making
2. Model Cards
3. Principles for Transparency
4. Ethics in Technology Practice

Markkula Center
for Applied Ethics
at Santa Clara University

II.1. The Markkula Framework for Ethical Decision Making

A comprehensive approach for making ethical decisions

Extremely general, useable for any case

Not a formula for a simple solution, but a process for

- Managing complexity
- Better understanding ethical problems
- Perceiving better choices
- Making better choices

Markkula Center
for Applied Ethics
at Santa Clara University

II.1. The Markkula Framework for Ethical Decision Making

1. Recognize the Ethical Issues: What values and risks are involved? Who are the stakeholders?
2. Get the Facts: What do we need to know? Who do we need to hear from?
3. Evaluate Alternative Actions through Multiple Ethical Lenses: What values do they prioritize? What harms & benefits will they bring? To whom?
4. Make a Decision and Mentally Test It: What's the ethical call, based on what we know? How would it hold up under scrutiny?
5. Act and Reflect on Outcomes: How did it turn out? What did we learn?

Markkula Center
for Applied Ethics
at Santa Clara University

II.1. The Markkula Framework for Ethical Decision Making

1. Recognize the Ethical Issues: What values and risks are involved? Who are the stakeholders?
2. Get the Facts: What do we need to know? Who do we need to hear from?
3. **Evaluate Alternative Actions through Multiple Ethical Lenses:** What values do they prioritize? What harms & benefits will they bring? To whom?
4. Make a Decision and Mentally Test It: What's the ethical call, based on what we know? How would it hold up under scrutiny?
5. Act and Reflect on Outcomes: How did it turn out? What did we learn?

Markkula Center
for Applied Ethics
at Santa Clara University

II.1.3. Evaluate Alternative Actions through Multiple Ethical Lenses

1. **The Utilitarian Approach**: Which option will produce the most good and do the least harm?
2. **The Rights Approach**: Which option best respects the rights of all who have a stake?
3. **The Justice Approach**: Which option treats people equally or proportionately?
4. **The Common Good Approach**: Which option best serves the community as a whole, not just some members?
5. **The Virtue Approach**: Which option leads me towards becoming a better person?
6. **The Care Approach:** Which option is the most caring thing to do?

Markkula Center
for Applied Ethics
at Santa Clara University

II.2. Model Cards

From a 2018/19 arXiv paper by Mitchell et al.
https://arxiv.org/abs/1810.03993

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phe-

KEYWORDS

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

ACM Reference Format:
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287596

1 INTRODUCTION

Markkula Center
for Applied Ethics
at Santa Clara University

II.2. Model Cards

- A model card acts something like a "nutrition label" for an AI model
- An approach to transparency for answering basic questions about a model's nature, purpose, and content
- Both ask for them and create them

Model Card

- **Model Details.** Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use.** Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors.** Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics.** Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data.** Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data.** May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

II.2. Model Card Examples

Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

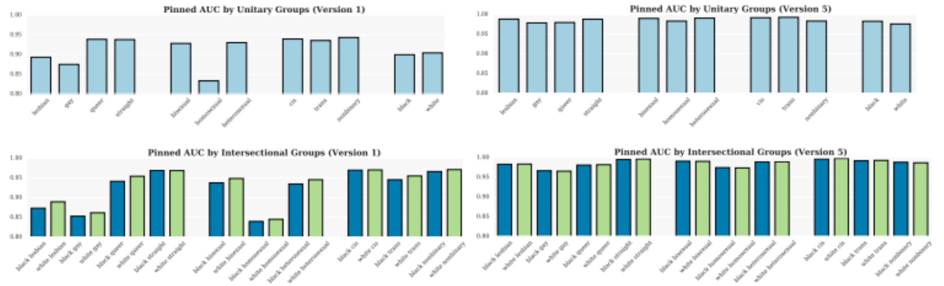
Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses



Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

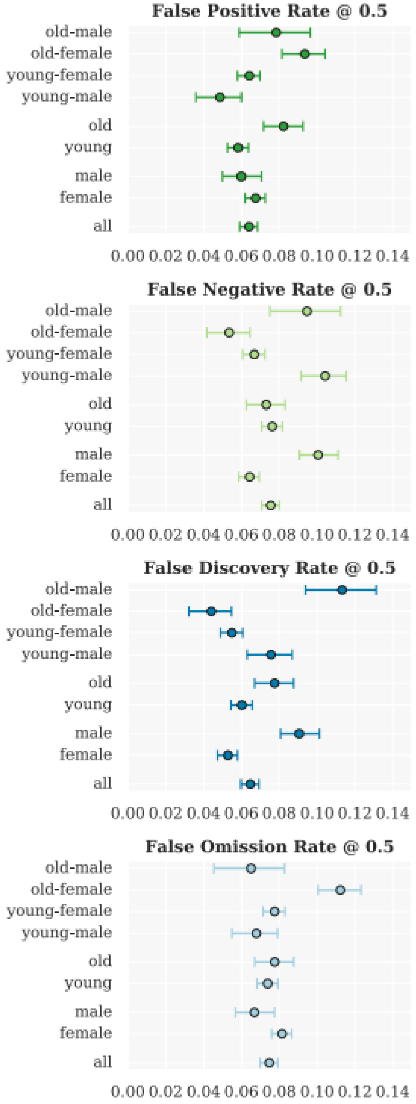
Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses



Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

II.3. ITEC – The Institute for Technology, Ethics, and Culture

A free resource for operationalizing tech ethics in organizations.

- A set of principles
- Stages for operationalizing principles
- A responsible technology management system

II.3. ITEC's Guiding Principles

1. Respect for Human Dignity and Rights
2. Promote Human Well-Being
3. Invest in Humanity
4. Promote Justice, Access, Diversity, Equity, and Inclusion
5. Recognize that Earth Is for All Life
6. Maintain Accountability
7. Promote Transparency and Explainability

Markkula Center
for Applied Ethics
at Santa Clara University

II.3. ITEC's Guiding Principles

1. Respect for Human Dignity and Rights
2. Promote Human Well-Being
3. Invest in Humanity
4. Promote Justice, Access, Diversity, Equity, and Inclusion
5. Recognize that Earth Is for All Life
6. Maintain Accountability
7. **Promote Transparency and Explainability**

Markkula Center
for Applied Ethics
at Santa Clara University

II.3. ITEC Transparency Principles (7, A-B)

7. Promote Transparency and Explainability – Accountability relies on being able to understand who and what made particular ethically significant choices and how and why those choices were made. Process… matters, and so the transparency and explainability of those processes matter too.

A. **Transparency & trustworthiness** – We commit to transparency with an aim to be considered a trustworthy enterprise. Trust comes from trustworthiness, and trustworthiness comes from a history of making the right choices for the right reasons…

B. **Simplicity** – products and services should be designed in the simplest way possible to reduce complexity…

II.3. ITEC Transparency Principles (C-E)

C. Fact-based decision-making – We commit to using facts. Decision making ought to be accountable to facts, not merely opinions or ideologies…

D. Openness on process and decision-making – We believe in openness in process and decision making. Closedness and secrecy harm trust. As much as possible, decision making ought to be open so that reasoning is visible and results are interpretable and accountable.

E. Human oversight – We value human oversight. All machine systems ought to have humans overseeing them so that there are people to appeal to for explanations, to prevent machine systems from going astray and causing harm, and to maintain accountability.

II.3. ITEC Transparency Principles (F-H)

F. Interpretability – We believe our products/services should be interpretable and understandable as well as the decisions from any human or machine system.

G. Reporting Status and Progress – We will report progress against a set of goals and identify the audiences they are serving in their decision making in a way that stakeholders can easily find and understand.

H. Feedback channels for explanations – We offer feedback channels for input and to provide explanations.

Markkula Center
for Applied Ethics
at Santa Clara University

II.4. Ethics in Technology Practice

- Piloted at Alphabet’s X “moonshot” division
- Materials being implemented and/or customized for several major companies, including Google; another for 60,000 employees
- Integratable into workflows and product design processes

Markkula Center
for Applied Ethics
at Santa Clara University

Ethics in Technology Practice

What Are These Materials? »

What We Need From You »

Overview of Ethics in Tech Practice

Conceptual Frameworks

Ethics in Technology Practice

- What Are These Materials?
- Overview of Ethics in Tech Practice
- Conceptual Frameworks
- Framework for Ethical Decision Making
- Ethical Toolkit
- Case Studies
- Sample Design Workflow
- Sample Workshop Slides
- Best Ethical Practices in Technology
- Authors

Framework for Ethical Decision Making

Ethics Toolkit

Case Studies

Sample Design Workflow

II.4. ETP's Ethical Toolkit

1. **Ethical Risk Sweeping:** Ethical risks are choices that may cause significant harm to persons or other entities with moral status.
2. **Ethical Pre-mortems and Post-mortems:** focuses on avoiding systemic ethical failures of a project.
3. **Expanding the Ethical Circle:** design teams need to invite stakeholder input and perspectives beyond their own.
4. **Case-based Analysis:** Case-based analysis enables ethical knowledge and skill transfer across ethical situations.
5. **Remembering the Ethical Benefits of Creative Work:** Ethical design and engineering is about human flourishing.
6. **Think About the Terrible People:** there will always be those who wish to abuse that power.
7. **Closing the Loop: Ethical Feedback and Iteration:** Ethical design and engineering is never a finished task

Markkula Center
for Applied Ethics
at Santa Clara University

Markkula Center for Applied Ethics at Santa Clara University

ITEC THE INSTITUTE FOR TECHNOLOGY, ETHICS, AND CULTURE

ETHICAL TOOLKIT

EXPANDING THE ETHICAL CIRCLE

Ensuring that the legitimate moral interests of all stake-holders have been taken into account, and that impacted communities have been consulted

ETHICAL PRE-MORTEMS

Exercising the skill of identify-ing how ethical failure of a project might happen and understanding the prevent-able causes so they can be mitigated

CASE-BASED ANALYSIS

Reviewing existing use cases with similar ethical dilemmas, to transfer knowledge and skill across ethical situations

ETHICAL RISK SWEEPING

Ethical risks are choices that may cause harm to persons or other entities with a moral status, or spark acute moral controversy. Failing to anticipate such risks can constitute ethical negligence. Ethical risk sweep-ing is an essential tool for good design and engineering practice.

ETHICAL POST-MORTEMS

Reviewing projects that fail, in order to identify the risks that were missed, the causes of ethical failure, what/who could have prevented it, and what can be done better next time

REMEMBERING ETHICAL BENEFITS

Keeping the ethical benefits at the center of the project, framing clearly its positive outcomes

THINKING ABOUT THE TERRIBLE PEOPLE

Identifying those groups or individuals who may abuse or misuse the technology and setting mitigation plans

CLOSING THE LOOP

Creating channels to invite ethically salient feedback, integrating with post-project data gathering and user support, and developing procedures for ethical iteration

Idea

- Technical and Market Research
- Casuistry (Case-Based Ethical Analysis of similar product ideas and their outcomes; what lessons/risks may transfer to this project?)

Sketch

- Early Storyboarding/Wireframing of Product Functionality
- Expanding the Circle (broaden your vision of stakeholder needs/capacities/interests beyond ideal user. Also, what are the dual-use cases?)

Feasability Assessment

- Technical Risk Scanning
- Ethical Risk Scanning (what are the most significant ethical risks/concerns that this project generates?
- Technical and Ethical Pre-Mortem (assume this project fails. Where will the *human/organizational* failure points be? How can we avoid them?)

Prototype

- Build
- Remember the Ethical Benefits of Creative Work (is this prototype still on track to deliver the envisioned benefits?)
- Expanding the Circle (is there anyone this prototype *won't* work for, or won't work as *well*? If so *who*, and why not? Can/must we fix that?)

Design (UX/UI)

- Code (Front End Experience)
- Expanding the Circle (*continued*: which users are we designing *for/not* designing for? Are we making the mistake of 'designing for ourselves'?)
- Think about the Terrible People (what openings/incentives might this code leave open for bad actors to exploit/abuse?)

Development

- Code (Back-End Functionality)
- (Keep) Thinking About the Terrible People (see above step; think again about dual-use contexts and whether your code invites harm there)
- Remember the Ethical Benefits Again (is this product *still* on track to deliver the envisioned benefit, or is it becoming something else?)

Testing

- UAT (User Acceptance) and Beta Testing
- Expanding the Circle (is our test group diverse/disinterested enough to reflect the real social impact of this tool? Or are we preaching to the choir?)
- Pre-Mortem (what test feedback might reveal an *ethical* failure point for this product in the future? Is there a bug fix here, or a deeper ethical flaw?)

Release

- Launch/Marketing
- Closing the Loop: Ethical Iteration (what feedback channels have we built to give us reliable signals of ethical concerns about the product?)
- (Keep) Thinking About The Terrible People (who may soon use this product in ways we didn't want, and what is our plan to limit their harm?)

Feedback

- Customer Support/Quality Management
- Expanding the Circle (are we soliciting *ethically* salient feedback from the *full* range of affected stakeholders (incl. non-users), or just a narrow set?)
- Closing The Loop (what are we *doing* with the ethically salient feedback we are getting? Who is responsible for analyzing/acting on it?)

Resources on the Markkula Center website

The Framework for Ethical Decision Making: https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/a-framework-for-ethical-decision-making/

The ITEC Handbook : https://www.scu.edu/institute-for-technology-ethics-and-culture/itec-handbook/

Ethics in Technology Practice: https://www.scu.edu/ethics-in-technology-practice/

Ethics in Technology Practice Toolkit: https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/

Markkula Center
for Applied Ethics
at Santa Clara University

Thank You!

Brian Patrick Green
Director of Technology Ethics
Markkula Center for Applied Ethics
Santa Clara University
bpgreen@scu.edu