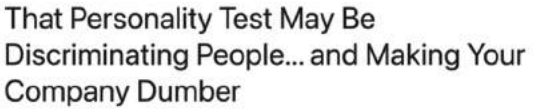
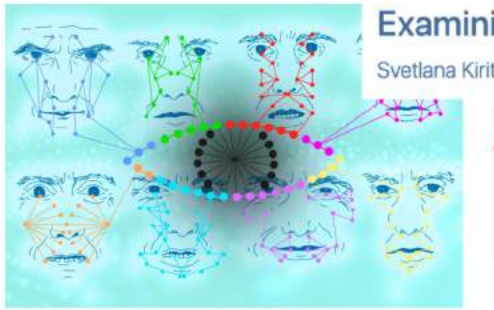
Ethics Sheets for AI Tasks

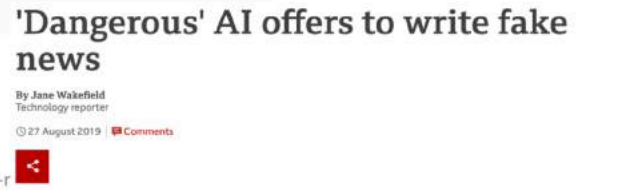
Saif M. Mohammad
Senior Research Scientist, National Research Council Canada
✉ Saif.Mohammad@nrc-cnrc.gc.ca 🐦 @SaifMMohammad

Technology Often at Odds with People

- more adverse outcomes for those that are already marginalized







Criticisms of AI Systems and Research

- Physiognomy, racism, bias, discrimination, perpetuating stereotypes, causing harm, ignoring indigenous world views, and more

- Thoughtlessness in machine learning
  - e.g., *is automating this task, this way, really going to help people?*

- Seemingly callous disregard for the variability and complexity of human behavior

(Fletcher-Watson et al. 2018; McQuillan 2018; Birhane 2021)

creativity,
diversity

What part do we play in this as researchers, system builders, leaders of tech companies?

What are the hidden assumptions in our research/work/product?

What are the unsaid implications of our choices?

Three Parts of this Work: Ethics Sheets for AI Tasks

- The Case
  - for documenting ethics considerations at the level of AI *Tasks*
- The Proposal
  - a new form of such an effort: *Ethics Sheets for AI Tasks*
- A Template and an Example
  - to facilitate creating new ethics sheets
  - example ethics sheet for *Automatic Emotion Recognition*

Recent Innovations to Bolster Ethics in AI/ML/NLP Research

- Individual Datasets: Datasheets (Gebru et al., 2018; Bender and Friedman, 2018)
- Individual Systems: Model cards (Mitchell et al., 2019)
- Individual Papers and Funding Applications: Broader Impacts Statements

Limitations:

- Conflict of interest for authors
  - strong incentives to present the work in positive light
- Tendency to produce boiler-plate text
- Scope is limited to individual pieces of work
- They are not meant to be public before the work begins

Need for engagement with ethics at a higher level

- beyond individual papers and add-on documents for individual projects

Important Ethical Considerations Apply at the Level of AI Tasks

Example Task: Detecting personality traits from one's history of utterances

Questions:

- What are the societal implications of automating personality trait detection?
- How can such a system be used/misused?
- What are the privacy implications of such a task?
- Is there enough credible scientific basis for personality trait identification?
- Which theory of personality traits should such automation rely on? What are the implications of that choice?
- What ethical considerations are latent in the choices we make in dataset creation, model development, and evaluation?

Not engaging with such questions have manifested in harms (and controversies for a number of AI tasks).

AI Tasks and Hazard (potential for harm / adverse effect)

- Face recognition
- Automatic Emotion Recognition (AER)
- Personality trait identification
- Machine translation
- Coreference resolution
- Image generation
- Text generation
- Text summarization
- Detecting trustworthiness
- Deception detection
- Information retrieval
- …

Female historians and male nurses do not exist, Google Translate tells its European users

by Nicolas Kayser-Bril

That Personality Test May Be Discriminating People... and Making Your Company Dumber

Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems

Svetlana Kiritchenko, Saif Mohammad

'Dangerous' AI offers to write fake news

By Jane Wakefield
Technology reporter

27 August 2019 | Comments

Should we create moral machines?

September 13, 2021 artificial intelligence, Center of Medical Ethics and Health Policy, Decision-making, Machine Ethics

AI Tasks and Hazard (potential for harm / adverse effect)

- Face recognition
- Automatic Emotion Recognition (AER)
- Personality trait identification
- Machine translation
- Coreference resolution
- Image generation
- Text generation
- Text summarization
- Detecting trustworthiness
- Deception detection
- Information retrieval
- …

All AI task technologies have associated:

- societal impact (varying degrees)
- **hazards**

Developing these technologies necessitates ethical considerations.

So:

If one wants to do work on an AI Task, it will be useful to have

a go-to point for the ethical considerations relevant to that task!

Especially since tens of thousands of new researchers and developers join the field every year.

This Talk: Ethics Sheets for AI Tasks

- The Case
  - for documenting ethics considerations at the level of *AI *Tasks**
- **The Proposal**
  - a new form of such an effort: *Ethics Sheets for AI Tasks*
- A Template and an Example
  - to create a sheet for your AI task
  - example ethics sheet for *Automatic Emotion Recognition*

A Call to Create Ethics Sheets for AI Tasks

Safety Data Sheets are required by law when dealing with hazardous materials

A document that substantively engages with the hazards and ethical issues relevant to a task:

- goes beyond individual systems and datasets
- draws on knowledge from a body of past work (from AI ethics, Psychology, Linguistics, Social Science, etc.)
  - like a survey article, but with a focus on ethics
- centers those most affected

Ethics Sheet for an AI Task

- Fleshes out assumptions
  - in how the task is commonly framed
  - in the choices often made regarding the data, method, and evaluation
- Presents hazards and ethical considerations unique / especially relevant to the task
- Communicates societal implications
  - to researchers, engineers, the broader public
- Lists common harm mitigation strategies

Not so much telling one what is right and wrong. More about helping one in their goal of determining what is appropriate in what context.

Target Audience

The various stakeholders of the AI Task:

- Researchers
- Engineers
- Educators (especially those who teach AI, ethics, or societal implications of technology)
- Media professionals
- Policy makers
- Politicians
- People whose data is used to train AI systems
- People impacted by AI systems
- Society at large

No One Sheet to Rule them All

A single ethics sheet does not speak for the whole community

Multiple ethics sheets

- by different teams and approaches
- reflect multiple perspectives and viewpoints

Record what is considered important to different groups of people at different times.

Components of an Ethics Sheet

- Preface
  - Why and how the sheet came to be written. The process. Who worked on it. Challenges faced. Changes (if a revision). Version, date published, contact info.
- Introduction
  - Task definition & terminology, scope, ways in which the task can manifest
- Motivations and Benefits
  - List of motivations, research interests, and commercial motivations of the task
- **Ethical Considerations**
  - A list of ethical considerations organized in groups; associated trade-offs, choices, societal implications, harm-mitigations strategies

This Talk: Ethics Sheets for AI Tasks

- The Case
  - for documenting ethics considerations at the level of *AI *Tasks**
- The Proposal
  - a new form of such an effort: *Ethics Sheets for AI Tasks*
- **A Template and an Example**
  - to create a sheet for your AI task
  - example ethics sheet for *Automatic Emotion Recognition*

Template

50 considerations grouped under:

- *Task Design*
- *Data*
- *Method*
- *Impact and Evaluation*

} common phases in system development

- *Implications for Privacy and Social Groups*

TASK DESIGN

A. Theoretical Foundations

1. Task Design and Framing
2. Theoretical Models and their Implications
3. Meaning and Extra-Linguistic Information
4. Wellness and Health Implications
5. Aggregate Level vs. Individual Level Prediction

B. Implications of Automation

6. Why Automate
7. Embracing Diversity
8. Participatory/Emancipatory Design
9. Applications, Dual Use, Misuse
10. Disclosure of Automation

DATA

C. Why This Data

11. Types of data
12. Dimensions of data

D. Human Variability–Machine Normativeness

13. Variability of Expression, Conceptualization
14. Norms of Emotions Expression
15. Norms of Attitudes
16. "Right" Label or Many Appropriate Ones…

Example: Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis

Medium Blog Post:
https://medium.com/@nlpscholar/ethics-sheet-aer-b8d671286682

To Appear in CL Journal June 2022

Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis

Saif M. Mohammad*

The importance and pervasiveness of emotions in our lives makes affective computing a tremendously important and vibrant line of work. Systems for automatic emotion recognition (AER) and sentiment analysis can be facilitators of enormous progress (e.g., in improving public health and commerce) but also enablers of great harm (e.g., for suppressing dissidents and manipulating voters). Thus, it is imperative that the affective computing community actively engage with the ethical ramifications of their creations. In this paper, I have synthesized and organized information from AI Ethics and Emotion Recognition literature to present fifty ethical considerations relevant to AER. Notably, the sheet fleshes out assumptions hidden in how AER is commonly framed, and in the choices often made regarding the data, method, and evaluation. Special attention is paid to the implications of AER on privacy and social groups. Along the way, key recommendations are made for responsible AER. The objective of the sheet is to facilitate and encourage more thoughtfulness on why to automate, how to automate, and how to judge success ***well before*** *the building of AER systems. Additionally, the sheet acts as a useful introductory document on emotion recognition (complementing survey articles).*

Key Questions:

- Is it even possible, or ethical, to determine one's internal mental state?
- What are the implications of human variability and creativity?
- Who is often left out in the design of existing systems?
- Which model of emotions is appropriate for a specific task/project?
- Are we carelessly endorsing questionable theories?
- Are AI systems conveying to the user what is "normal"; implicitly invalidating other forms of emotion expression?

NRC·CNRC

Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis

CL Journal, June 2022

PREFACE

Automatic Emotion Recognition (AER) can be a force that helps unlock:
- how emotions work; how they relate to our health, language, social interactions
- numerous commercial applications

Yet, AER can also be a tool for substantial harm:
- mass application on vulnerable populations
- unreliable approaches; privacy concerns; physiognomy

Should we be building AER systems? Are they ethical?

This sheet helps in thinking about these questions. It:
- documents and organizes ethical considerations
- discusses factors at play in particular contexts

Saif M. Mohammad
National Research Council Canada
http://saifmohammad.com
saif.mohammad@nrc-cnrc.gc.ca
@SaifMMohammad

No One Sheet to Rule them All

A single ethics sheet does not speak for the whole community

Multiple ethics sheets (by different teams, approaches) for the same or overlapping tasks can reflect multiple perspectives, viewpoints, and what is important to different groups of people at different times.

This sheet for AER is an example of "Ethics Sheets for AI Tasks" (ACL 2022)

A Call to Document Ethics Considerations at the Level of AI *Tasks*

INTRODUCTION

Scope: AER from text (AER in NLP)

Task: AER is an umbrella term for numerous tasks; e.g., inferring…
1. emotions felt by the speaker
2. emotions perceived by the listener
3. patterns of emotions over time
4. speaker's stance to a target
5. and many more…

Tasks & Modalities come with benefits, harms, ethical considerations

Modalities for AER
- facial expressions, gait, proprioceptive data (movement of body), gestures
- skin and blood conductance, blood flow, respiration, infrared emanations
- force of touch, haptic data
- speech, text

50 ETHICAL CONSIDERATIONS

I. TASK DESIGN

A. Theoretical Foundations
1. Emotion Task and Framing
2. Emotion Models and Choice of Emotions
3. Meaning, Extra-Linguistic Information
4. Wellness and Health Implications
5. Aggregate vs. Individual Level

B. Implications of Automation
6. Why Automate
7. Embracing Diversity
8. Participatory Design
9. Applications, Dual Use
10. Disclosure of Automation

II. DATA

C. Why This Data
11. Types of data
12. Dimensions of data

D. Human Variability v Machine Normativeness
13. Variability of Expression, Representation
14. Norms of Emotions Expression
15. Norms of Attitudes
16. "Right" Label or Many Appropriate Ones
17. Label Aggregation
18. Historical Data
19. Training-Deployment Differences

E. The People Behind the Data
20. Platform Terms of Service
21. Anonymization and Deletion
22. Warnings and Recourse
23. Crowdsourcing

III. METHOD

F. Why This Method
24. Types of Methods and Tradeoffs
25. Who is Left Out by this Method
26. Spurious Correlations
27. Context is Everything
28. Individual Emotion Dynamics
29. Historical Behavior
30. Emotion Management, Manipulation
31. Green AI

IV. IMPACT AND EVALUATION

G. Metrics
32. Reliability/Accuracy
33. Demographic Biases
34. Sensitive Applications
35. Testing (Diverse Datasets, Metrics)

H. Beyond Metrics
36. Interpretability, Explainability
37. Visualization
38. Safeguards and Guard Rails
39. Harms when System Works as Designed
40. Contestability and Recourse
41. Be wary of Ethics Washing

V. PRIVACY, SOCIAL GROUPS

I. Implications for Privacy
42. Privacy and Personal Control
43. Group Privacy and Soft Biometrics
44. Mass Surveillance vs. Right to Privacy, Expression, Protest
45. Right Against Self-Incrimination
46. Right to Non-Discrimination

J. Implications for Social Groups
47. Disaggregation
48. Intersectionality
49. Reification and Essentialization
50. Attributing People to Social Groups

What are the ethical considerations for your task?

1. Emotion Task and Framing

Is the goal to infer one's emotions from an utterance?
- is it possible to do so?
- is it ethical to try to infer such a personal mental state?

Often, other framings are more appropriate.

2. Emotion Model and Choice of Emotions

Avoid careless endorsement of discredited ideas:
- universality of some emotions; basic emotions
- universal mapping to facial expressions (Barrett 2017)
- internal state related to outward appearance: physiognomy

8. Participatory/Emancipatory Design

Paper

Poster

"nothing about us without us"
- disabilities research (Stone and Priestley 1996)
- indigenous communities research (Hall 2014)

Center people, especially disadvantaged communities (Oliver 1997; Spinuzzi 2005, Noel 2016)
- agency to shape the design process

13-19. Human Variability v Machine Normativeness

variability in mental representation, expression of emotions
vs.
inherent bias of modern machine learning approaches to focus on what is common (in the training data)

Through their behaviour (e.g., recognizing some forms of expressions and not others), AI systems convey to the user what is "normal"; implicitly invalidating other forms.

43. Group Privacy

Soft-biometrics
- identifying groups of people with similar traits
- people disfavour such profiling (McStay, 2020)

There are very few Moby-Dicks. Most of us are sardines. The individual sardine may believe that the encircling net is trying to catch it. It is not. It is trying to catch the whole shoal. It is therefore the shoal that needs to be protected, if the sardine is to be saved. — Floridi (2014)

National Research Council Canada | Conseil national de recherches Canada

Canada

Ethics Sheet for AER and Sentiment Analysis

- Poster: Tomorrow 11am, 7pm
- Talk: Wednesday 1:30pm

  (sentiment track)

NRC·CNRC

Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis

CL Journal, June 2022

PREFACE

Automatic Emotion Recognition (AER) can be a force that helps unlock:
- how emotions work; how they relate to our health, language, social interactions
- numerous commercial applications

Yet, AER can also be a tool for substantial harm:
- mass application on vulnerable populations
- unreliable approaches; privacy concerns; physiognomy

Should we be building AER systems? Are they ethical?

This sheet helps in thinking about these questions. It:
- documents and organizes ethical considerations
- discusses factors at play in particular contexts

Saif M. Mohammad
National Research Council Canada
http://saifmohammad.com
saif.mohammad@nrc-cnrc.gc.ca
@SaifMMohammad

No One Sheet to Rule them All

A single ethics sheet does not speak for the whole community

Multiple ethics sheets (by different teams, approaches) for the same or overlapping tasks can reflect multiple perspectives, viewpoints, and what is important to different groups of people at different times.

This sheet for AER is an example of "Ethics Sheets for AI Tasks" (ACL 2022)

A Call to Document Ethics Considerations at the Level of AI *Tasks*

INTRODUCTION

Scope: AER from text (AER in NLP)

Task: AER is an umbrella term for numerous tasks; e.g., inferring…
1. emotions felt by the speaker
2. emotions perceived by the listener
3. patterns of emotions over time
4. speaker's stance to a target
5. and many more…

Tasks & Modalities come with benefits, harms, ethical considerations

Modalities for AER
- facial expressions, gait, proprioceptive data (movement of body), gestures
- skin and blood conductance, blood flow, respiration, infrared emanations
- force of touch, haptic data
- speech, **text**

50 ETHICAL CONSIDERATIONS

I. TASK DESIGN

A. Theoretical Foundations
1. Emotion Task and Framing
2. Emotion Models and Choice of Emotions
3. Meaning, Extra-Linguistic Information
4. Wellness and Health Implications
5. Aggregate vs. Individual Level

B. Implications of Automation
6. Why Automate
7. Embracing Diversity
8. Participatory Design
9. Applications, Dual Use
10. Disclosure of Automation

II. DATA

C. Why This Data
11. Types of data
12. Dimensions of data

D. Human Variability v Machine Normativeness
13. Variability of Expression, Representation
14. Norms of Emotions Expression
15. Norms of Attitudes
16. "Right" Label or Many Appropriate Ones
17. Label Aggregation
18. Historical Data
19. Training-Deployment Differences

E. The People Behind the Data
20. Platform Terms of Service
21. Anonymization and Deletion
22. Warnings and Recourse
23. Crowdsourcing

III. METHOD

F. Why This Method
24. Types of Methods and Tradeoffs
25. Who is Left Out by this Method
26. Spurious Correlations
27. Context is Everything
28. Individual Emotion Dynamics
29. Historical Behavior
30. Emotion Management, Manipulation
31. Green AI

IV. IMPACT AND EVALUATION

G. Metrics
32. Reliability/Accuracy
33. Demographic Biases
34. Sensitive Applications
35. Testing (Diverse Datasets, Metrics)

H. Beyond Metrics
36. Interpretability, Explainability
37. Visualization
38. Safeguards and Guard Rails
39. Harms when System Works as Designed
40. Contestability and Recourse
41. Be wary of Ethics Washing

V. PRIVACY, SOCIAL GROUPS

I. Implications for Privacy
42. Privacy and Personal Control
43. Group Privacy and Soft Biometrics
44. Mass Surveillance vs. Right to Privacy, Expression, Protest
45. Right Against Self-Incrimination
46. Right to Non-Discrimination

J. Implications for Social Groups
47. Disaggregation
48. Intersectionality
49. Reification and Essentialization
50. Attributing People to Social Groups

What are the ethical considerations for your task?

1. Emotion Task and Framing

Is the goal to infer one's emotions from an utterance?
- is it possible to do so?
- is it ethical to try to infer such a personal mental state?

Often, other framings are more appropriate.

2. Emotion Model and Choice of Emotions

Avoid careless endorsement of discredited ideas:
- universality of some emotions; basic emotions
- universal mapping to facial expressions (Barrett 2017)
- internal state related to outward appearance: physiognomy

8. Participatory/Emancipatory Design

Paper

Poster

"nothing about us without us"
- disabilities research (Stone and Priestley 1996)
- indigenous communities research (Hall 2014)

Center people, especially disadvantaged communities (Oliver 1997; Spinuzzi 2005, Noel 2016)
- agency to shape the design process

13-19. Human Variability v Machine Normativeness

variability in mental representation, expression of emotions
vs.
inherent bias of modern machine learning approaches
to focus on what is common (in the training data)

Through their behaviour (e.g., recognizing some forms of expressions and not others), AI systems convey to the user what is "normal"; implicitly invalidating other forms.

43. Group Privacy

Soft-biometrics
- identifying groups of people with similar traits
- people disfavour such profiling (McStay, 2020)

There are very few Moby-Dicks. Most of us are sardines. The individual sardine may believe that the encircling net is trying to catch it. It is not. It is trying to catch the whole shoal. It is therefore the shoal that needs to be protected, if the sardine is to be saved. — Floridi (2014)

National Research Council Canada | Conseil national de recherches Canada

Canada

Ethics Sheets for AI Tasks: Discussion

Other relevant questions…

- Should we create ethics sheets for a handful of AI Tasks (more prone to being misused, say) or for all AI tasks?
- Who should create an Ethics Sheet for a AI task?
- Does it matter what we define as a `task'?
- Why Should Academic Researchers Care about this?
- How can we further incentivize researchers to create Ethics Sheets?
- When should we be creating Ethics Sheets for AI Tasks?
- Should we think about research systems differently from deployed systems?
- Is there a time dimension for these ethics sheets?
- How should we use ethics sheets?

Benefits of Ethics Sheets for AI Tasks

1. Encourages more thoughtfulness regarding why to automate, how to automate, and how to judge success
2. Fleshes out assumptions hidden in how the task is commonly framed, and in the choices often made regarding data, method, and evaluation
3. Presents the trade-offs of relevant choices so that stakeholders can make informed decisions appropriate for their context
4. Identifies points of agreement and disagreement
5. Moves us towards consensus and community standards
6. Helps us better navigate research and implementation choices
7. Has citations and pointers

Benefits of Ethics Sheets for AI Tasks (continued)

8. Helps stakeholders challenge assumptions
9. Helps all stakeholders develop harm mitigation strategies
10. Standardized sections and a familiar look and feel make it easy for the compilation and communication of ethical considerations
11. Helps in developing better datasheets and model cards
12. Engages the various stakeholders of an AI task with each other
13. Multiple ethics sheets reflect multiple perspectives, viewpoints, and what is considered important to different groups of people at different times
14. Acts as a great introductory document for an AI Task (complements survey articles and task-description papers for shared tasks).

Notable Benefits of an Ethics Sheet

Encourages more thoughtfulness regarding why to automate, how to automate, and how to judge success

A written document allows stakeholders to challenge our assumptions and conclusions …and that is a good thing!

Acts as a record of what we, as a community, value

Great introductory reading for a topic …complementing survey articles.

Poster

- In person: 5 to 6 pm today (session 3)
- Virtual: 7:30 to 8:30 am tomorrow

Take Home Message

- document ethics considerations at the **task level**

What are the ethical considerations for your task?

Slides, Proposal, Paper, Ethics Sheet for Emotion Recognition
Available at: www.saifmohammad.com

saif.Mohammad@nrc-cnrc.gc.ca
@SaifMMohammad