Ethics Sheets for AI Tasks

And a Case Study for Automatic Emotion Recognition

Saif M. Mohammad

Senior Research Scientist, National Research Council Canada

✉ Saif.Mohammad@nrc-cnrc.gc.ca   @SaifMMohammad

National Research Council Canada   Conseil national de recherches Canada

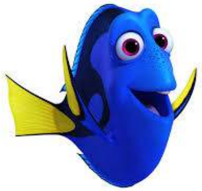
Canada

First, a brief overview of my general research interest...

Emotions

- Determine human experience
- Condition our actions
- Central in organizing meaning

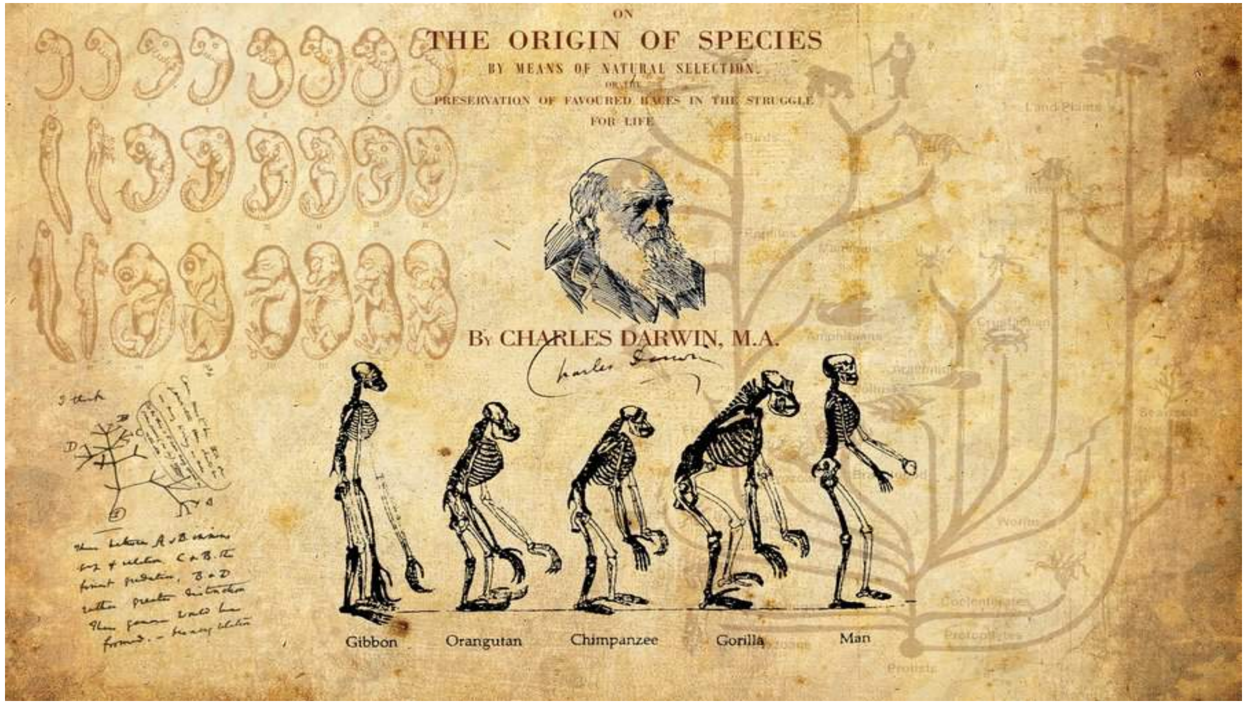
The Search for Emotions in Language

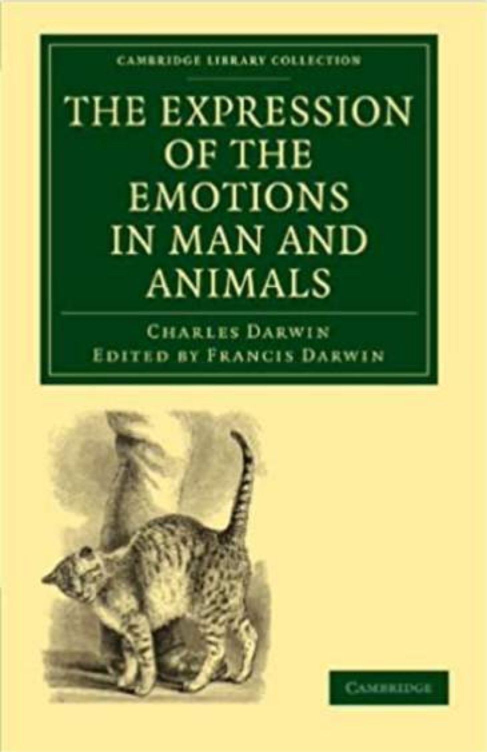


creativity

fairness

Study of Emotions and Psychological Models of Emotions

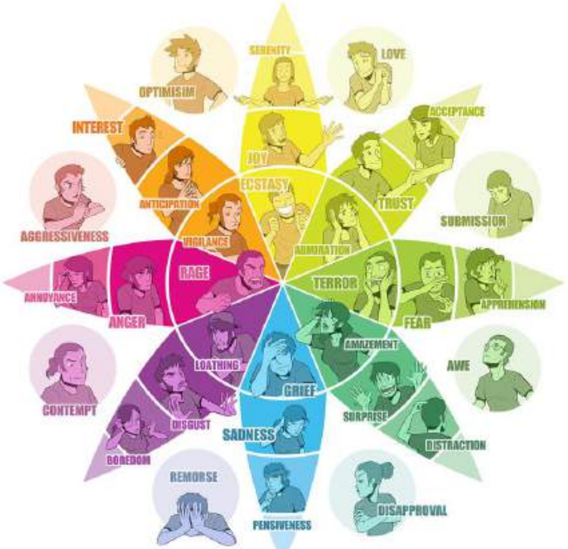



CAMBRIDGE LIBRARY COLLECTION
THE EXPRESSION OF THE EMOTIONS IN MAN AND ANIMALS
CHARLES DARWIN
EDITED BY FRANCIS DARWIN
CAMBRIDGE

Basic Emotions Theory (BET)

Some categorical emotions (joy, sadness, fear, etc.) are more basic than others

- Paul Ekman, 1971: Six Basic Emotions
- Plutchik, 1980: Eight Basic Emotions
- And many others



Plutchik's Emotion Wheel

Image credit: Julia Belyanevych

Some important tenets of BET discredited (Barrett, 2018)

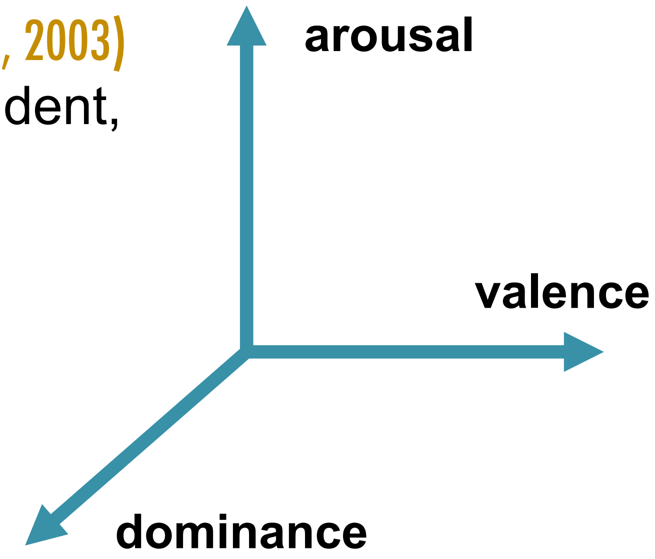
- Still useful to work on categorical emotions
  - How people talk about emotions

Theory of Constructed Emotion (Barrett, 2017)

Dimensional Theory of Emotions

Influential factor analysis studies (Osgood et al., 1957; Russell, 1980, 2003) have shown that the three most important, largely independent, dimensions of word meaning:

- valence (V): positive/pleasure – negative/displeasure
- arousal (A): active/stimulated – sluggish/bored
- dominance (D): powerful/strong – powerless/weak



Thus, when comparing the meanings of two words, we can compare their V, A, D scores. For example:

- *banquet* indicates more positiveness than *funeral*
- *nervous* indicates more arousal than *lazy*
- *queen* indicates more dominance than *delicate*

Word-Emotion Association Lexicons

Over the years, we have created lexicons for both categorical emotions as well as for valence, arousal, and dominance

- Lists of words associated with joy, sadness, fear, etc.
- Lists of words and their valence, arousal, and dominance scores

Through Manual annotations

- Comparative annotations (not Likert scales)
  - Avoids various biases

Automatically:

- Corpus-based methods

The NRC Valence, Arousal, and Dominance Lexicon (2018)

provides ratings of valence, arousal, and dominance for ~20,000 English words

http://saifmohammad.com/WebPages/nrc-vad.html

The NRC Word–Emotion Association Lexicon aka NRC Emotion Lexicon or EmoLex (2010)

provides associations for ~14,000 words with eight emotions (anger, fear, joy, sadness, anticipation, disgust, surprise, trust)

http://saifmohammad.com/WebPages/NRC-Emotion- Lexicon.htm

The NRC Emotion Intensity Lexicon aka Affect Intensity Lexicon (2018-19)

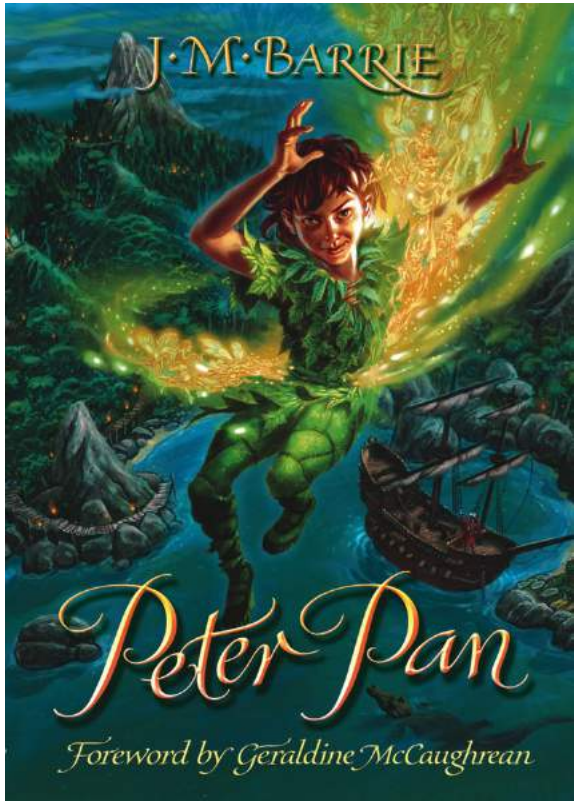
provides intensity scores for ~6000 words with four emotions (anger, fear, joy, sadness)

http://saifmohammad.com/WebPages/AffectIntensity.htm

The NRC Word–Colour Association Lexicon (2010)

provides associations for ~14,000 words with 11 common colours

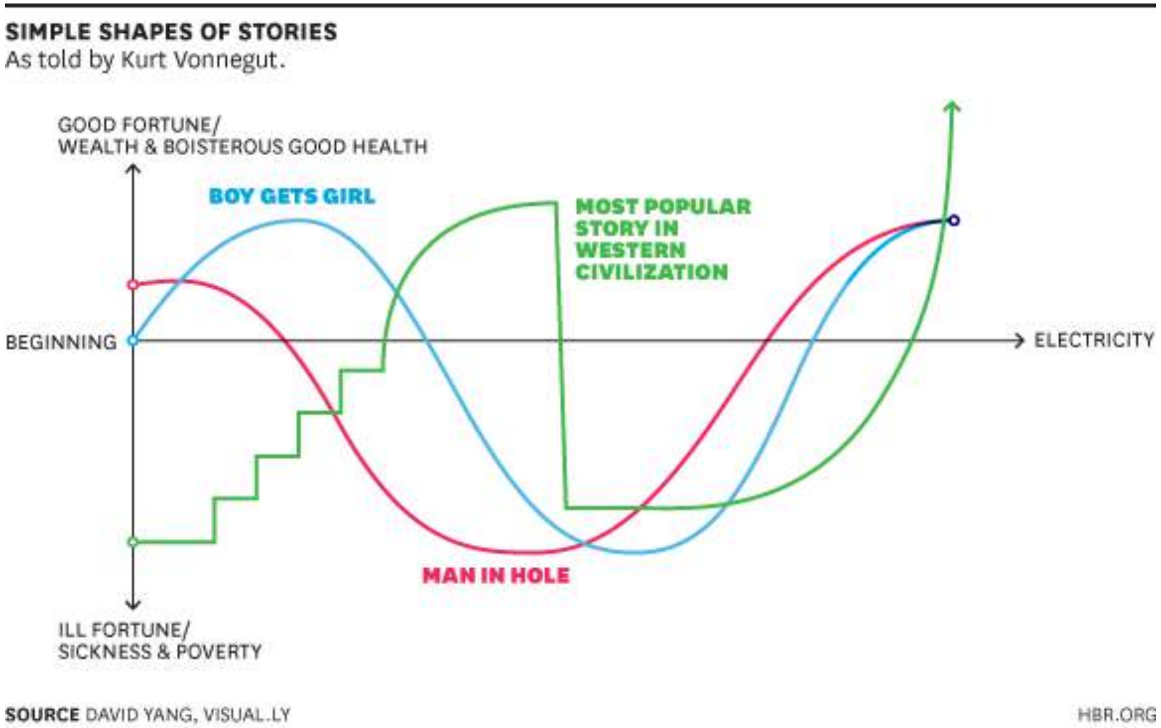
http://saifmohammad.com/WebPages/lexicons.html creativity



Detecting Emotions in Stories

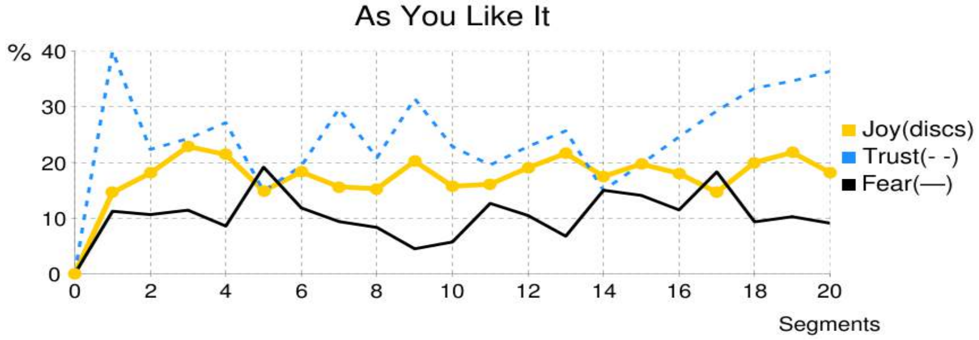
Tracking Emotions in Stories (Kurt Vonnegut inspired)

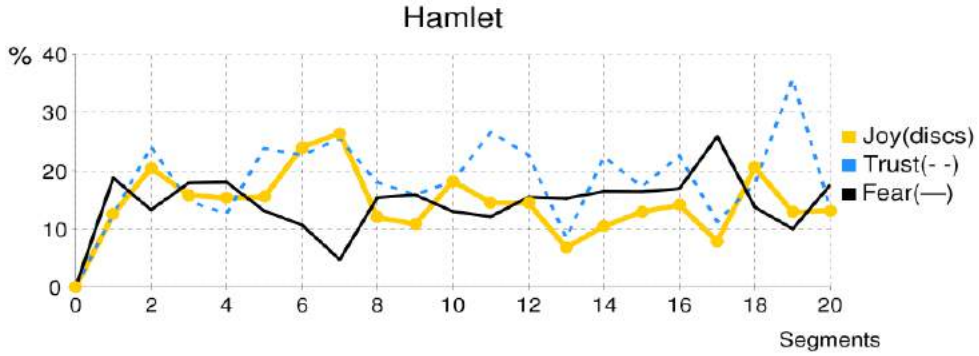
- Can we automatically track the emotions of characters?
- Are there some canonical shapes common to most stories?
- Can we track the change in distribution of emotion words?

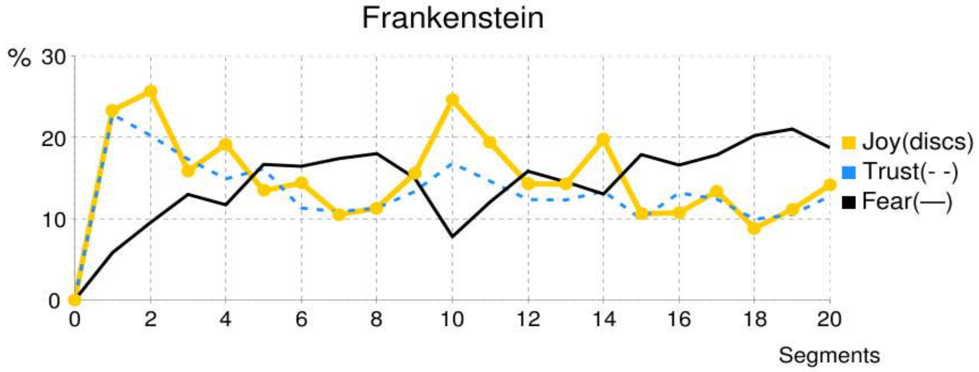


Back in 2011:

Tracking emotion word distribution in novels and fairy tales.







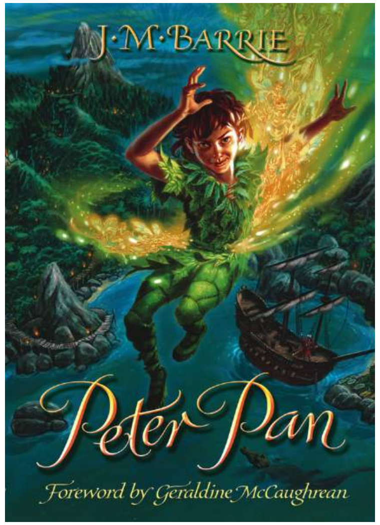
From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales, Saif Mohammad, In Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), June 2011, Portland, OR.

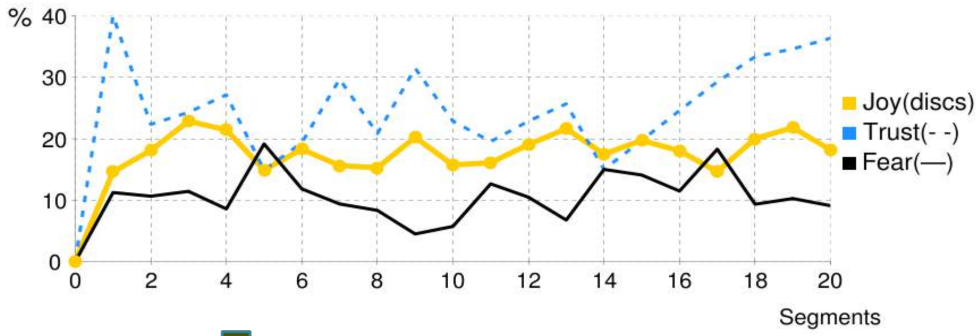
Hannah Davis
Artist/Programmer

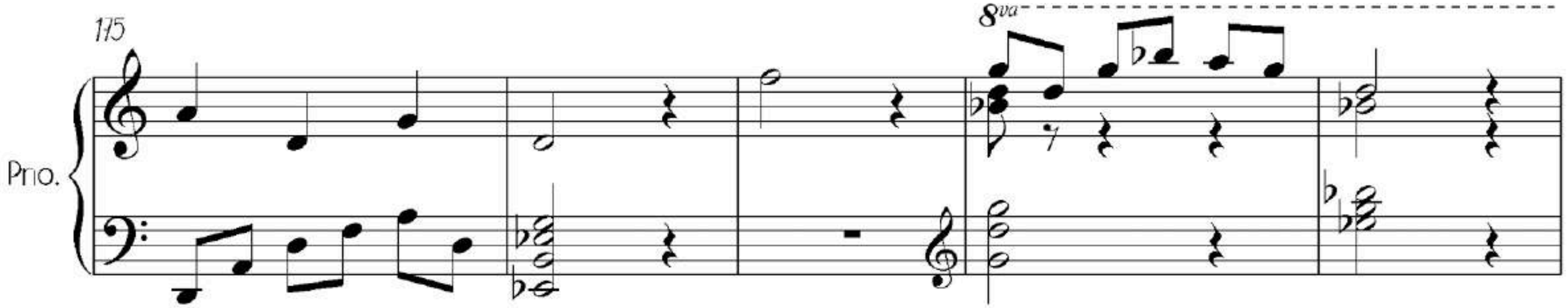
Generating music from text

Paper:

- Generating Music from Literature. Hannah Davis and Saif M. Mohammad, In Proceedings of the EACL Workshop on Computational Linguistics for Literature, April 2014, Gothenburg, Sweden.







A method to generate music from literature

- music that captures the change in the distribution of emotion words

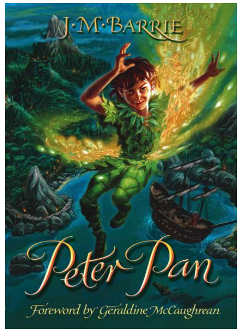
TransProse

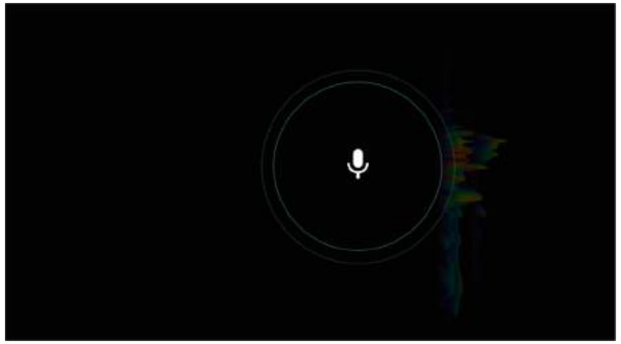
Automatically generates three simultaneous piano melodies pertaining to the dominant emotions in the text, using the NRC Emotion Lexicon.

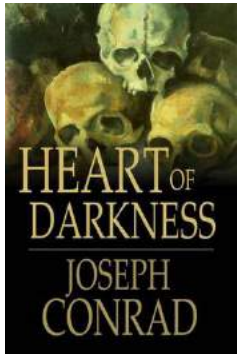
TransProse

Automatically generates three simultaneous piano melodies pertaining to the dominant emotions in the text, using the NRC Emotion Lexicon.

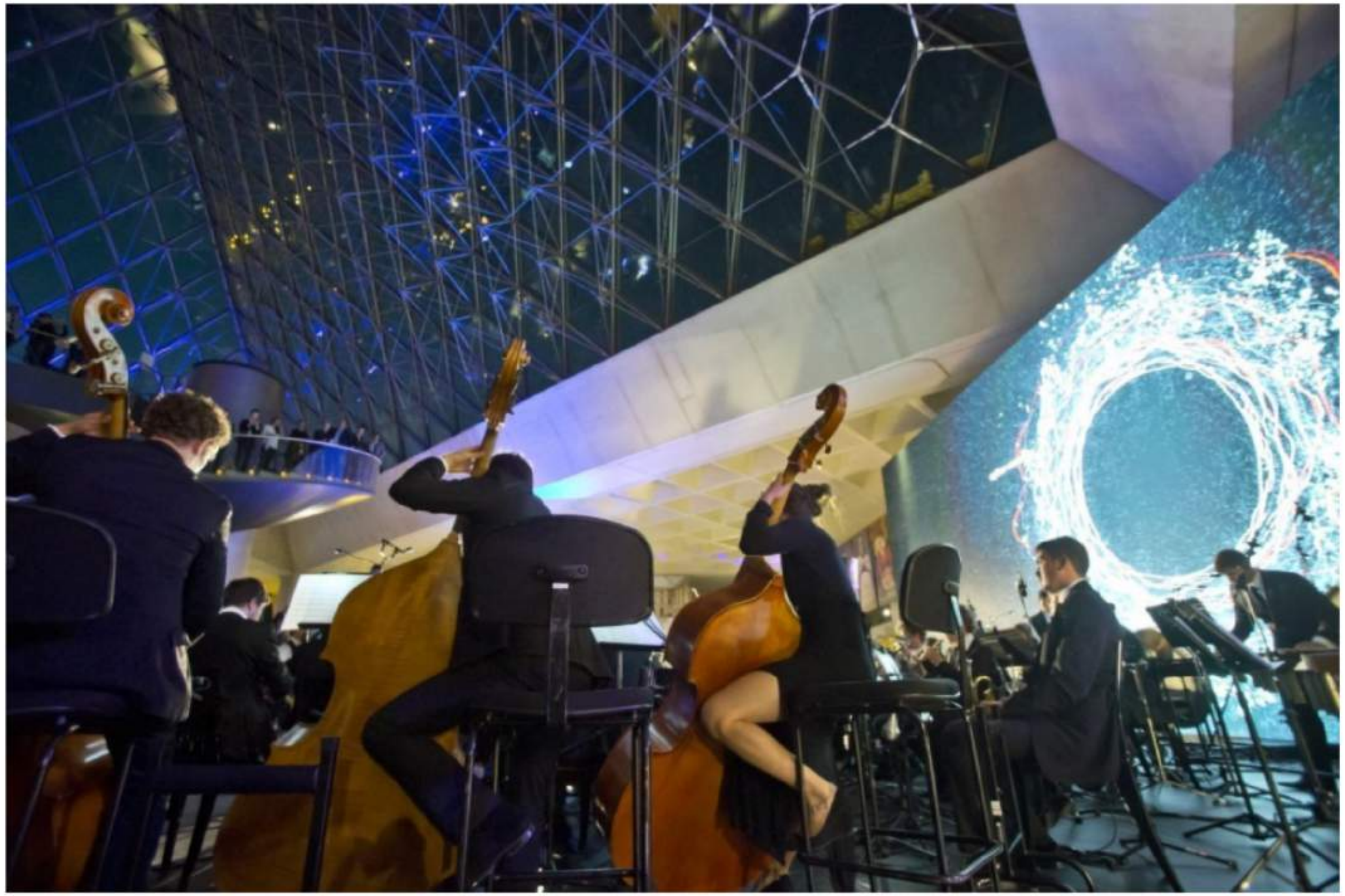
Examples



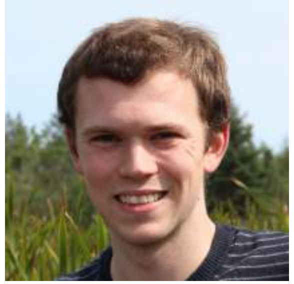




TransProse Music Played by an Orchestra, at the Louvre Museum, Paris



A symphony orchestra performs under the glass of the Louvre museum in Paris on Sept. 20. Accenture Strategy has created a symphonic experience enabled by human insight and artificial intelligence technology. (Michel Euler/AP)



Will Hipson

Psychologist/Programmer

(Picture by: Mulyadi)

Emotion Dynamics of Fictional Characters

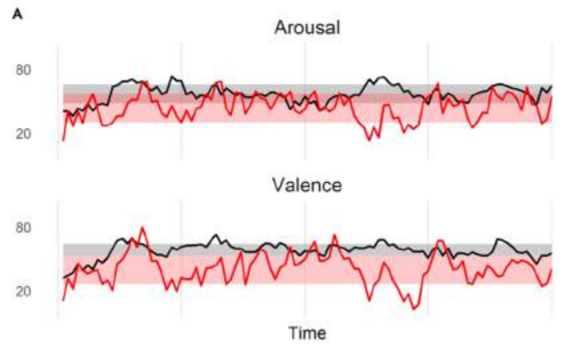
Paper:

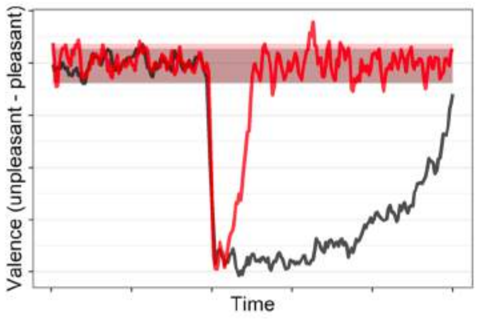
- Emotion Dynamics in Movie Dialogues. Will E. Hipson, Saif M. Mohammad, Under Review. (Paper available on ArXiv.)

Emotion Dynamics (from Psychology)

Study of change in emotional state with time

- intensive longitudinal data (repeated self-reports of emotional state)
- quite difficult to obtain such data





Another window into emotions is through our words:

- E.g., if happier, we are likely to utter more happiness-associated words

Utterance Emotion Dynamics: study of change in emotion words over time (Hipson and Mohammad, 2021)

Felipe José Bravo Márquez

Parinaz Sobhani

Shared Tasks on Emotions

Mohammad Salameh

Svetlana Kiritchenko

- Participation
  - Sentiment in Twitter (SemEval-2013, SemEval-2014)
    - 1st out of 40+ teams in multiple sub-tasks
- Organization
  - Detecting Stance (SemEval-2016)
  - Multiple emotion and sentiment tasks at word and tweet-level

Xiaodan Zhu

Colin Cherry

SemEval-2018 Task 1: Affect in Tweets

https://competitions.codalab.org/competitions/17751

Tasks: Inferring likely affectual state the tweeter is trying to convey

- emotion intensity regression and ordinal classification
- multi-label emotion classification task

Multiple emotion dimensions

English, Arabic, and Spanish Tweets

75 Team (~200 participants)

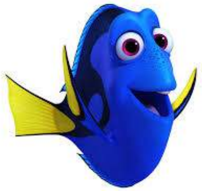
fairness

Included a separate evaluation component for biases towards race and gender.

Do systems give higher emotion intensity scores to mentions of one race/gender.

Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. Svetlana Kiritchenko and Saif M. Mohammad. In Proceedings of *Sem, New Orleans, LA, USA, June 2018.

The Search for Emotions in Language (continues)



creativity

fairness

Ethics Sheets for AI Tasks

And a Case Study for Automatic Emotion Recognition

National Research Council Canada
Conseil national de recherches Canada

Canada

This Talk

- The Case
  - Make a case for documenting ethics considerations at the level of *AI *Tasks**
- The Proposal
  - Propose a new form of such an effort: *Ethics Sheets for AI Tasks*
- The Example
  - Provide an example ethics sheet for *Automatic Emotion Recognition*

Discussion and FAQ

Target audience: AI, ML, NLP researchers and developers; educators;
all stakeholders of AI applications

Abbreviations: Artificial Intelligence (AI), Machine learning (ML), Natural Language Processing (NLP)

Paper: Ethics Sheets for AI Tasks. Saif M. Mohammad. *arXiv preprint arXiv:*2107.01183*.* July 2021.

Feedback: Welcome! The new proposal can benefit from more ideas.

Contact:

✉ Saif.Mohammad@nrc-cnrc.gc.ca 🐦 @SaifMMohammad

- **The Case**
  - Make a case for documenting ethics considerations at the level of *AI *Tasks**

- The Proposal
  - Propose a new form of such an effort: *Ethics Sheets for AI Tasks*

- The Example
  - Provide an example ethics sheet for *Automatic Emotion Recognition*

Discussion and FAQ

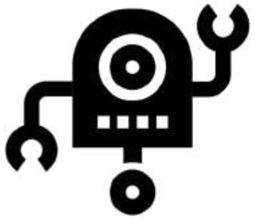
Good design helps everyone.

Why Designing for Accessibility Helps Everyone

Written by Daniel Abrahams

As NLP and ML systems become more ubiquitous, their broad societal impacts are receiving more scrutiny than ever before.

- technology is often at odds with the people
- more adverse outcomes for those that are already marginalized



Created by Oksana Latysheva from Noun Project

Do Machines Make Fair Decisions?

What part do we play in this as researchers?

What are the hidden assumptions in our research?

What are the unsaid implications of our choices?

Are we perpetuating and amplifying inequities
or are we striking at the barriers to opportunity?

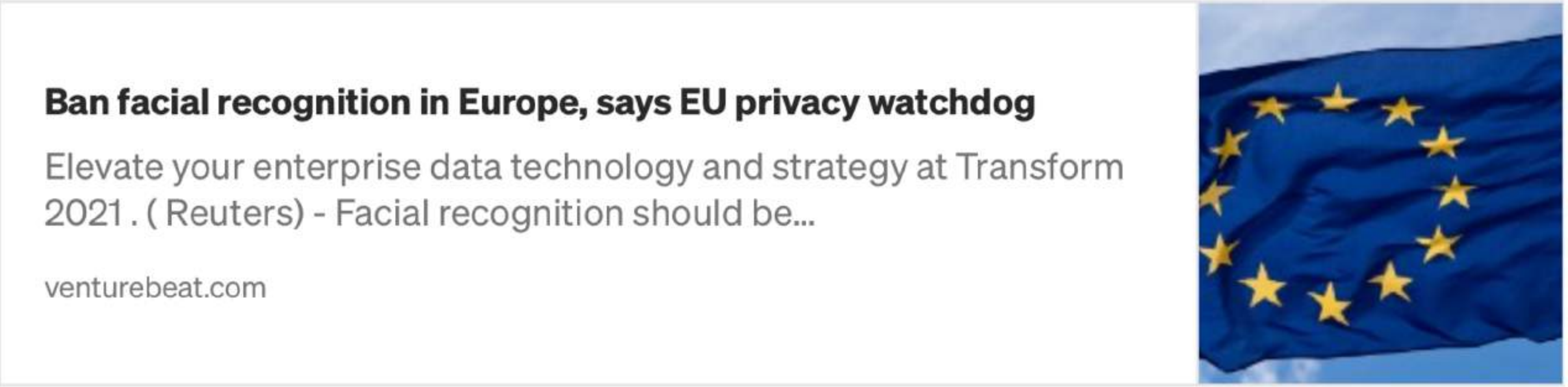
Answers are often complex and multifaceted.

Do People Make Fair Decisions?

Examples of Real-World Systems Gone Wrong

Created by Oksana Latysheva from Noun Project

- Microsoft's racist chatbot, Tay, posts inflammatory and racist tweets
- Amazon's AI recruiting tool biased against women
- Face recognition systems good for detecting faces of light-skinned men, but really bad for dark-skinned women
- Recidivism systems biased against people from African American neighborhoods
- Mass surveillance of vulnerable populations by governments



Ban facial recognition in Europe, says EU privacy watchdog

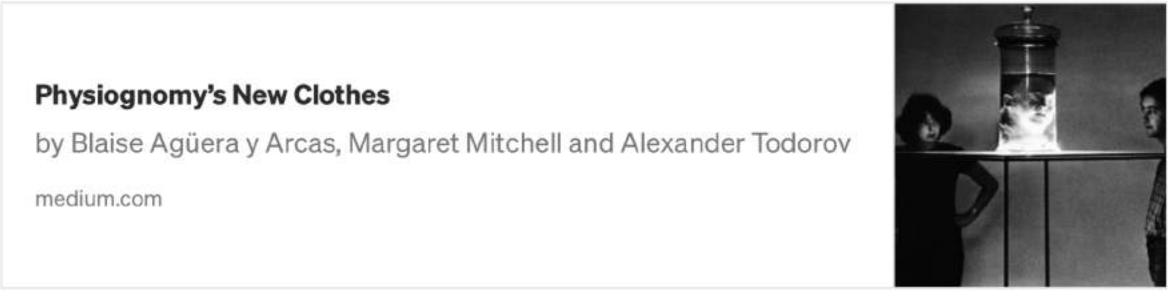
Elevate your enterprise data technology and strategy at Transform 2021 . (Reuters) - Facial recognition should be...

venturebeat.com

Criticisms of Published Research

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

Arcas, Mitchell, and Todorov (2017):

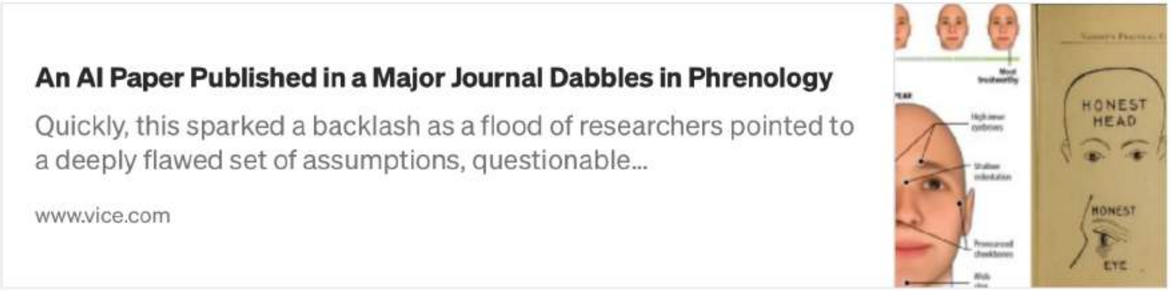


Physiognomy's New Clothes

by Blaise Agüera y Arcas, Margaret Mitchell and Alexander Todorov

medium.com

Ongweso Jr (2020):



An AI Paper Published in a Major Journal Dabbles in Phrenology

Quickly, this sparked a backlash as a flood of researchers pointed to a deeply flawed set of assumptions, questionable...

www.vice.com

Criticisms of Published 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)

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

- Individual Datasets
  - Datasheets: list key details of the datasets such as composition and intended use
- Individual Systems
  - Model cards: list key details of the models such as performance in various contexts and intended use scenarios
- Individual Papers
  - Ethics and impact statements, ethics reviews

Datasheets and Model cards

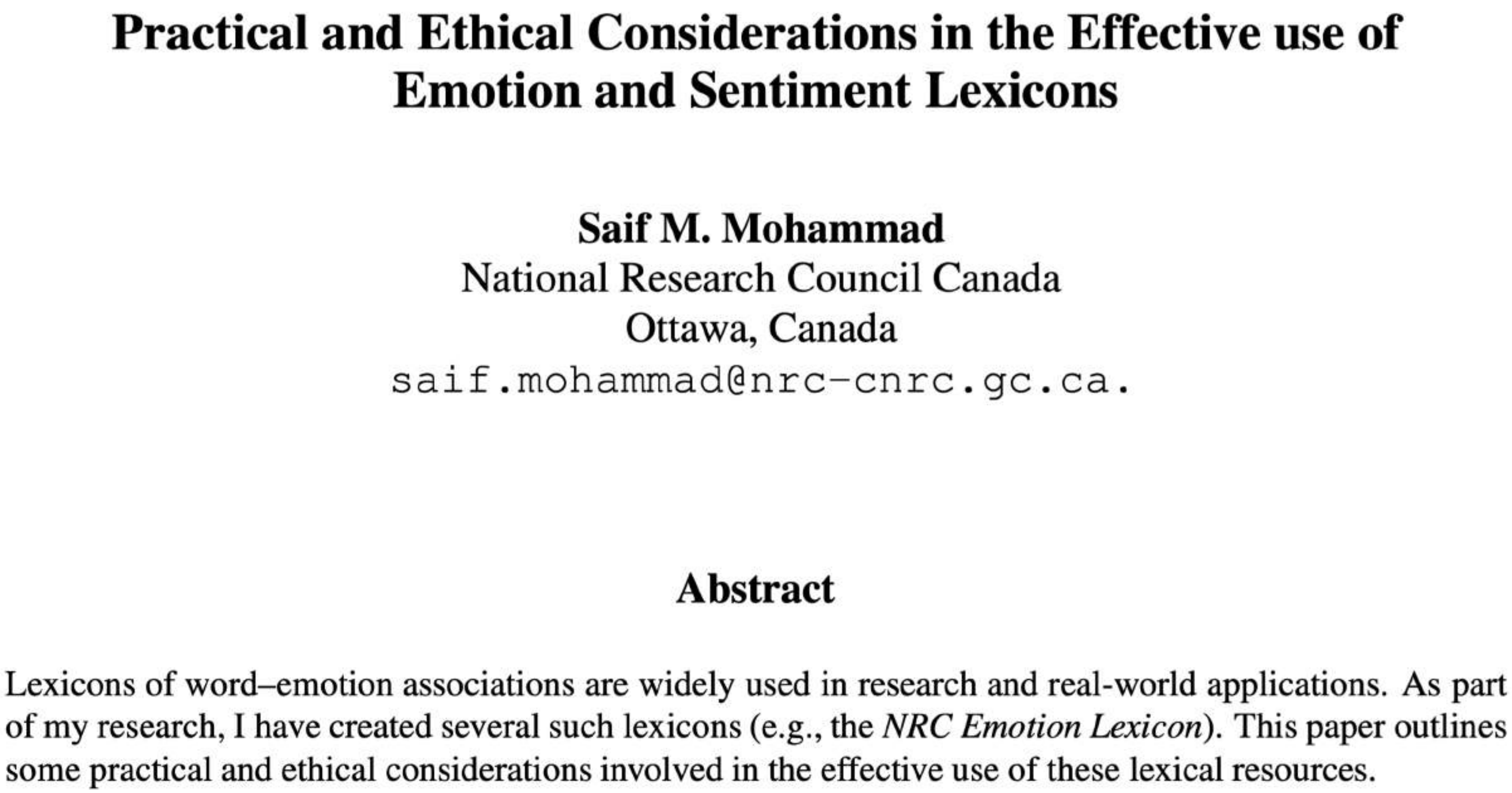
Pivotal inventions; but have limitations, notably due to scope

- 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

There is a need for engagement at a level beyond individual papers and add-on documents for individual projects:

- ethics considerations apply at various levels, including:
    - for whole areas of work
    - for AI tasks

Papers Looking at Specific Areas of Research



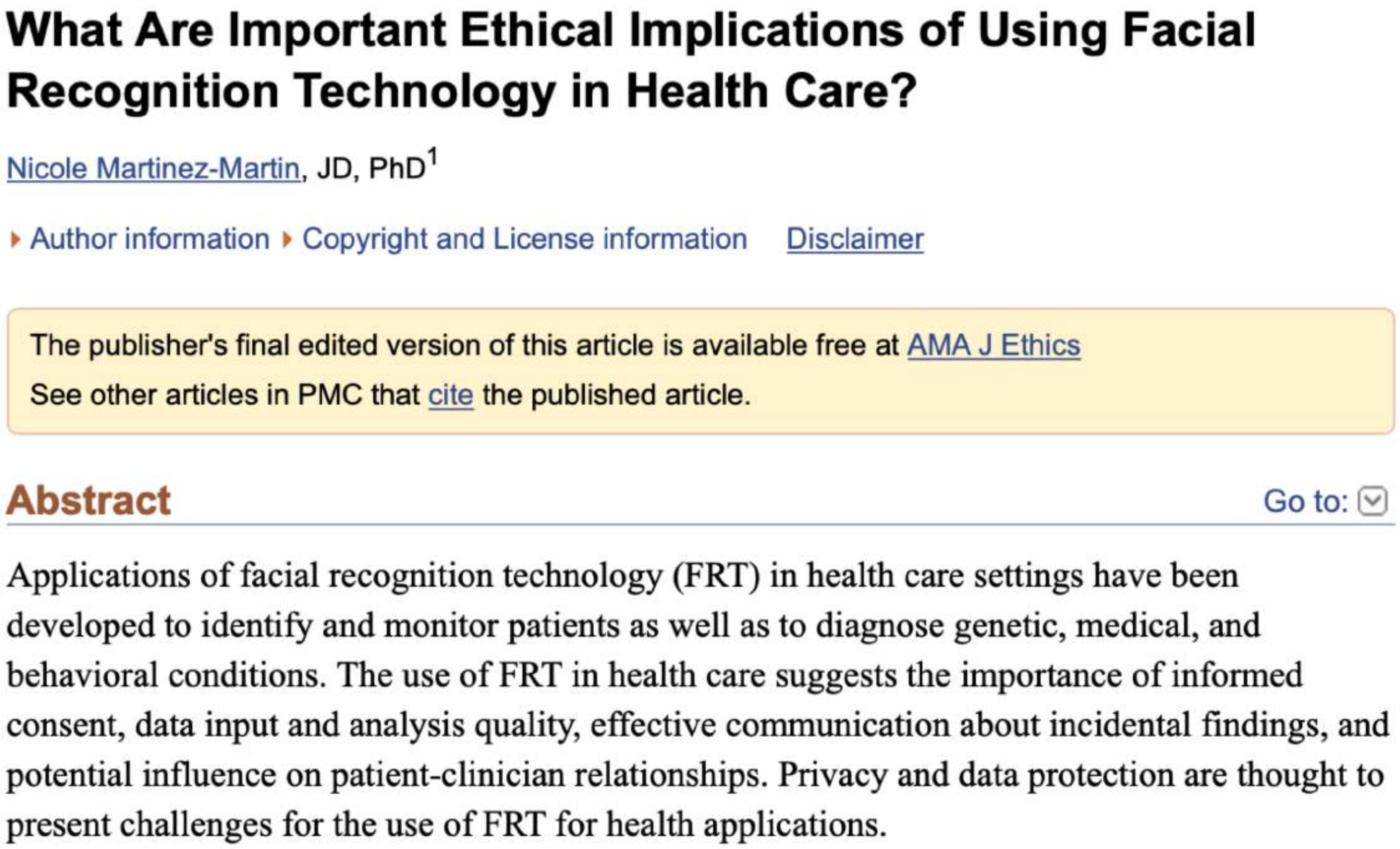
Practical and Ethical Considerations in the Effective use of Emotion and Sentiment Lexicons

Saif M. Mohammad
National Research Council Canada
Ottawa, Canada
saif.mohammad@nrc-cnrc.gc.ca.

Abstract

Lexicons of word–emotion associations are widely used in research and real-world applications. As part of my research, I have created several such lexicons (e.g., the *NRC Emotion Lexicon*). This paper outlines some practical and ethical considerations involved in the effective use of these lexical resources.

Papers Looking at Specific Areas of Research



What Are Important Ethical Implications of Using Facial Recognition Technology in Health Care?

Nicole Martinez-Martin, JD, PhD[1]

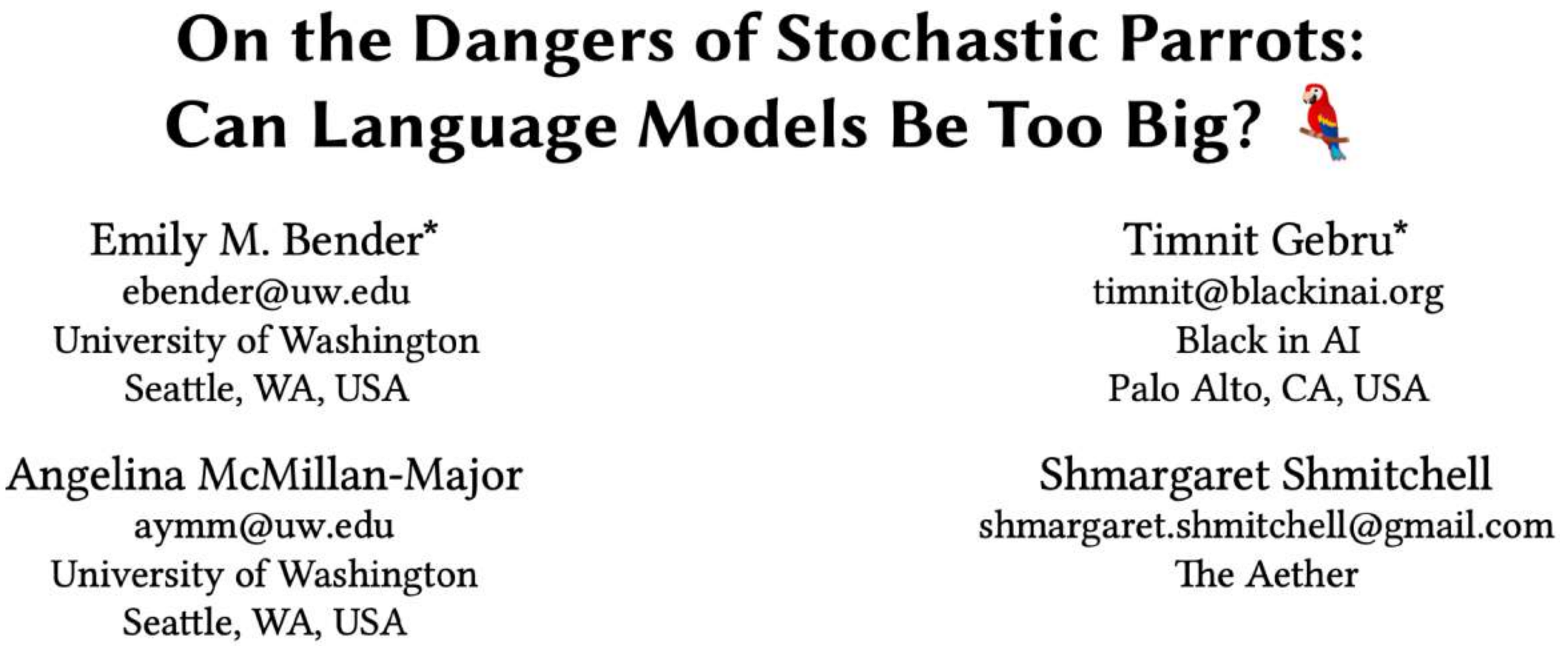
The publisher's final edited version of this article is available free at AMA J Ethics

See other articles in PMC that cite the published article.

Abstract

Applications of facial recognition technology (FRT) in health care settings have been developed to identify and monitor patients as well as to diagnose genetic, medical, and behavioral conditions. The use of FRT in health care suggests the importance of informed consent, data input and analysis quality, effective communication about incidental findings, and potential influence on patient-clinician relationships. Privacy and data protection are thought to present challenges for the use of FRT for health applications.

Papers Looking at Specific Areas of Research



On the Dangers of Stochastic Parrots:
Can Language Models Be Too Big? 🦜

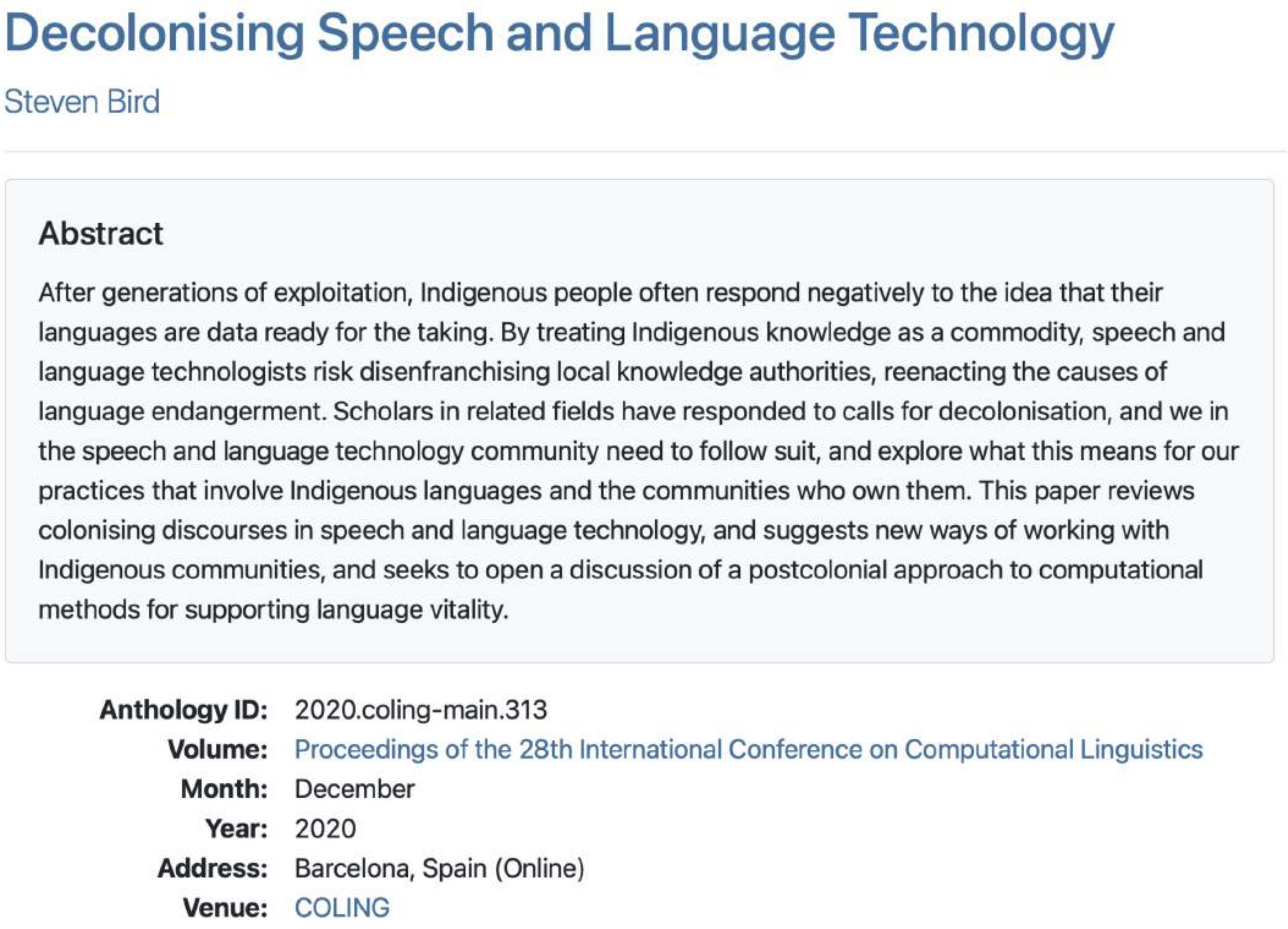
Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

Papers Looking at Specific Areas of Research



Decolonising Speech and Language Technology

Steven Bird

Abstract

After generations of exploitation, Indigenous people often respond negatively to the idea that their languages are data ready for the taking. By treating Indigenous knowledge as a commodity, speech and language technologists risk disenfranchising local knowledge authorities, reenacting the causes of language endangerment. Scholars in related fields have responded to calls for decolonisation, and we in the speech and language technology community need to follow suit, and explore what this means for our practices that involve Indigenous languages and the communities who own them. This paper reviews colonising discourses in speech and language technology, and suggests new ways of working with Indigenous communities, and seeks to open a discussion of a postcolonial approach to computational methods for supporting language vitality.

Anthology ID:	2020.coling-main.313
Volume:	Proceedings of the 28th International Conference on Computational Linguistics
Month:	December
Year:	2020
Address:	Barcelona, Spain (Online)
Venue:	COLING

Ethical Considerations Also Apply at the Level of AI Tasks

Tasks in AI Research

AI Task: some task we may want to automate using AI techniques

AI System: is a particular AI model built to do the task

Research in machine Learning:

- tasks used as test beds for evaluating methods

Research in Natural Language Processing, Computer Visions, Human-Computer Interaction, etc.:

- strong focus on tasks as a way to organize research/sub-fields

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 that we should attempt to do this?
- Which theory of personality traits should such automation rely on? What are the implications of that choice? and so on.

Currently, AI conferences and journals do not have a dedicated place for such a discussion.

Tasks in AI Research

In addition:

- ethical considerations are latent in the choices we make in dataset creation, model development, and evaluation

Poor choices have manifested in harms and controversies for a number of AI tasks.

So:

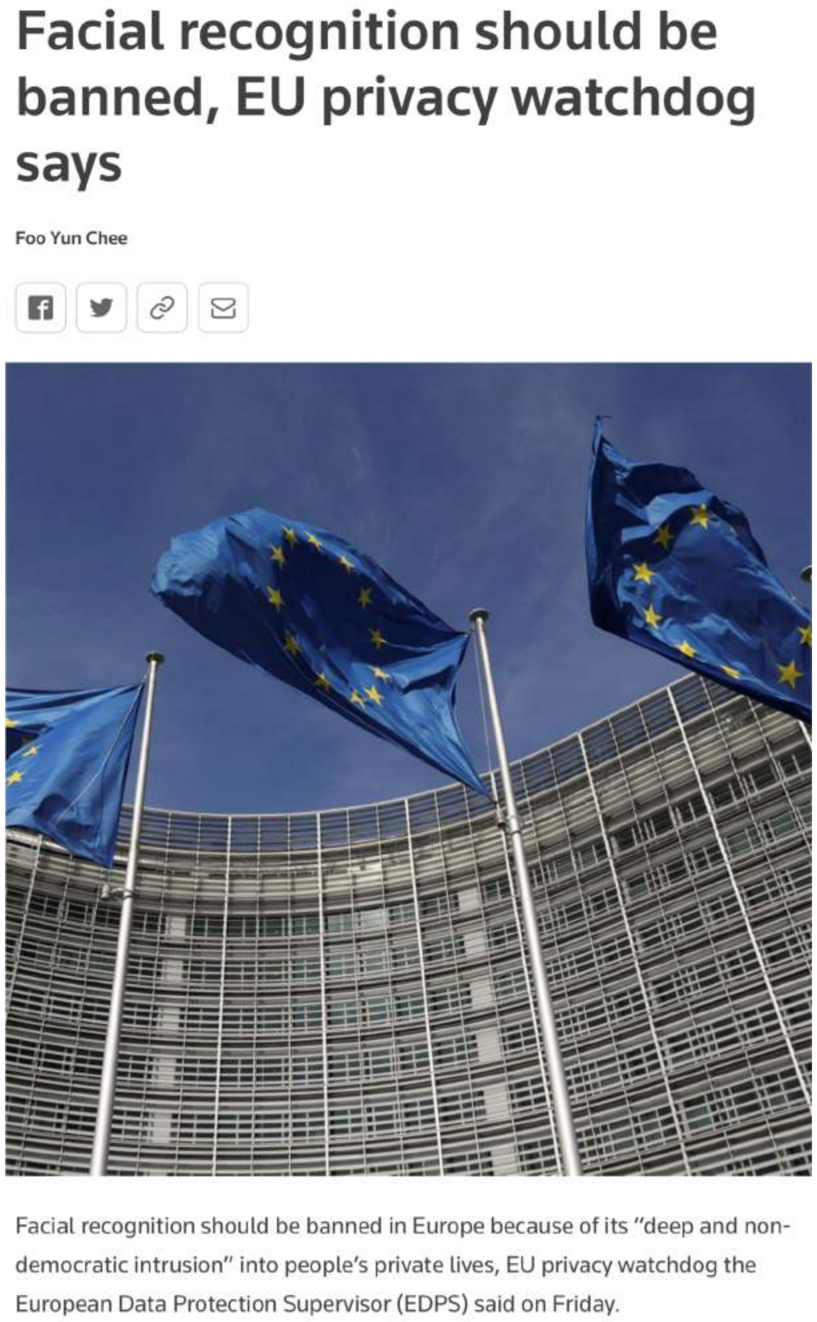
> If one wants to do work on an AI Task, it will be useful to have a go-to point for the ethical considerations relevant to that task!

but...

Do we need ethics sheets for tasks other than face recognition?

AI Tasks & Controversy

- Face recognition



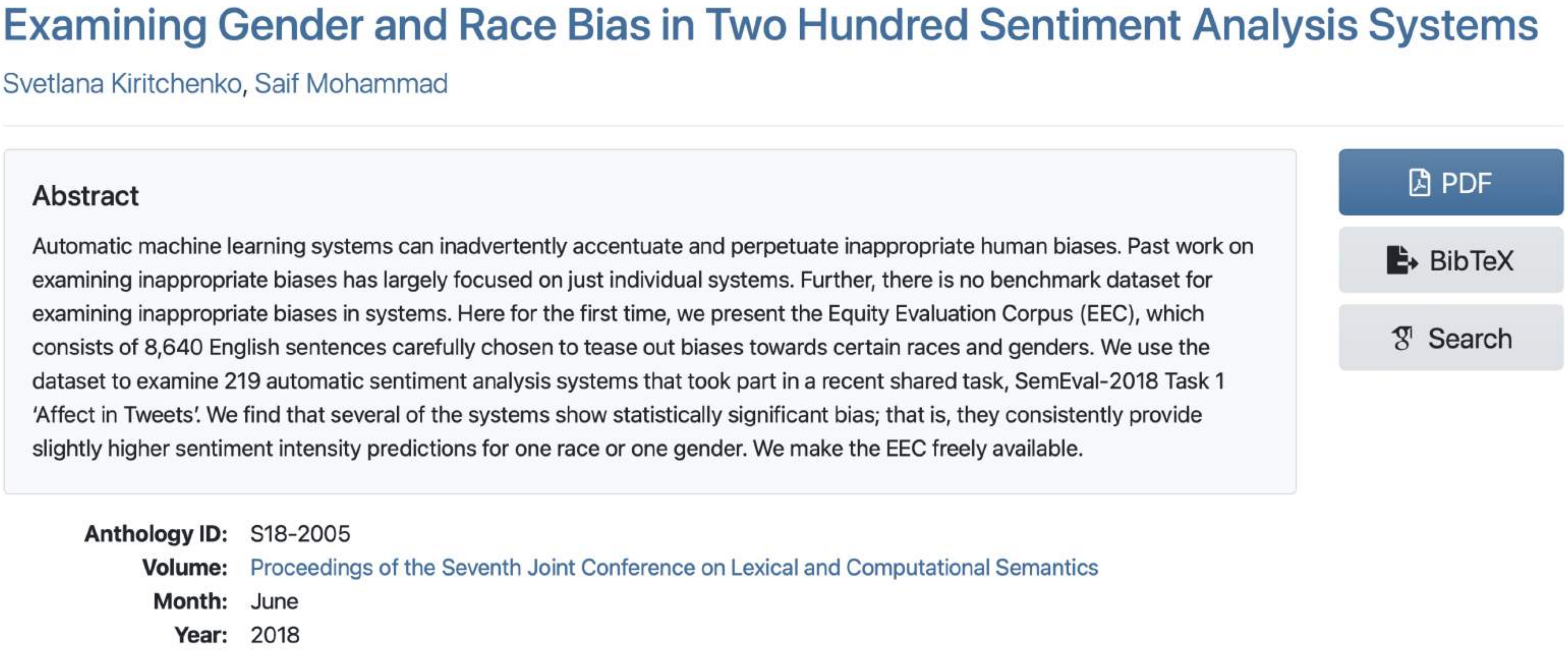
Facial recognition should be banned, EU privacy watchdog says

Foo Yun Chee

Facial recognition should be banned in Europe because of its "deep and non-democratic intrusion" into people's private lives, EU privacy watchdog the European Data Protection Supervisor (EDPS) said on Friday.

AI Tasks & Controversy

- Face recognition
- Emotion recognition



Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems

Svetlana Kiritchenko, Saif Mohammad

Abstract

Automatic machine learning systems can inadvertently accentuate and perpetuate inappropriate human biases. Past work on examining inappropriate biases has largely focused on just individual systems. Further, there is no benchmark dataset for examining inappropriate biases in systems. Here for the first time, we present the Equity Evaluation Corpus (EEC), which consists of 8,640 English sentences carefully chosen to tease out biases towards certain races and genders. We use the dataset to examine 219 automatic sentiment analysis systems that took part in a recent shared task, SemEval-2018 Task 1 'Affect in Tweets'. We find that several of the systems show statistically significant bias; that is, they consistently provide slightly higher sentiment intensity predictions for one race or one gender. We make the EEC freely available.

Anthology ID: S18-2005
Volume: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics
Month: June
Year: 2018

AI Tasks & Controversy

- Face recognition
- Emotion recognition

(unholy) combination

POLICY

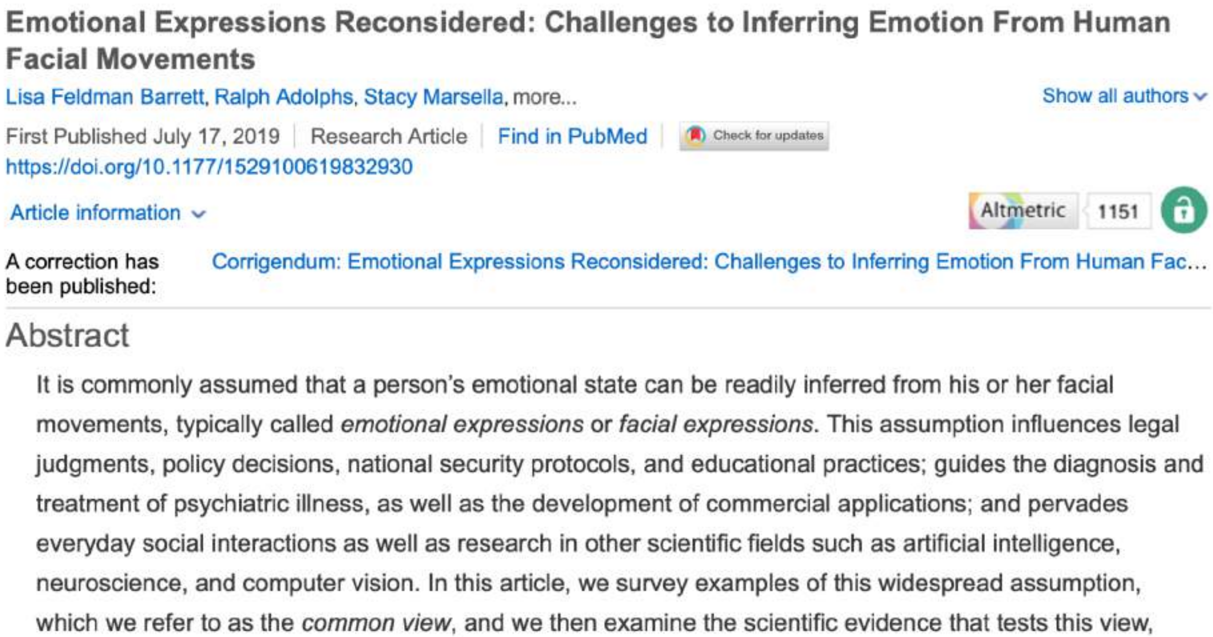
AI 'EMOTION RECOGNITION' CAN'T BE TRUSTED

The belief that facial expressions reliably correspond to emotions is unfounded, says a new review of the field

By James Vincent | Jul 25, 2019, 11:55am EDT

As artificial intelligence is used to make more decisions about our lives, engineers have sought out ways to make it more emotionally intelligent. That means automating some of the emotional tasks that come naturally to humans — most notably, looking at a person's face and knowing how they feel.

To achieve this, tech companies like Microsoft, IBM, and Amazon all sell what they call "emotion recognition" algorithms, which infer how people feel based on facial analysis. For example, if someone has a furrowed brow and pursed lips, it means they're angry. If their eyes are wide, their eyebrows are raised, and their mouth is stretched, it means they're afraid, and so on.



Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, more...

First Published July 17, 2019 | Research Article | Find in PubMed

https://doi.org/10.1177/1529100619832930

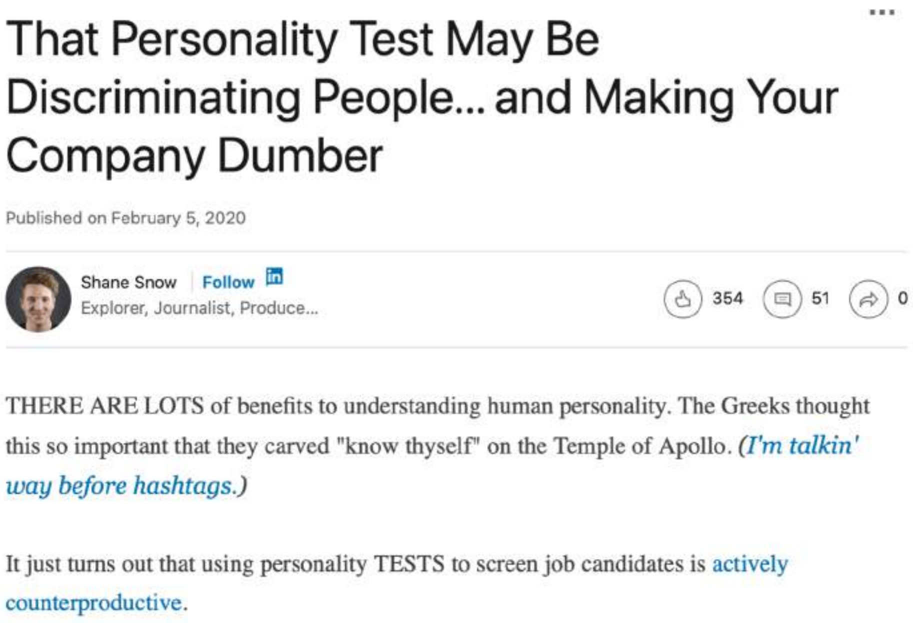
A correction has been published: Corrigendum: Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Fac...

Abstract

It is commonly assumed that a person's emotional state can be readily inferred from his or her facial movements, typically called *emotional expressions* or *facial expressions*. This assumption influences legal judgments, policy decisions, national security protocols, and educational practices; guides the diagnosis and treatment of psychiatric illness, as well as the development of commercial applications; and pervades everyday social interactions as well as research in other scientific fields such as artificial intelligence, neuroscience, and computer vision. In this article, we survey examples of this widespread assumption, which we refer to as the *common view*, and we then examine the scientific evidence that tests this view,

AI Tasks & Controversy

- Face recognition
- Emotion recognition
- Personality trait identification



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

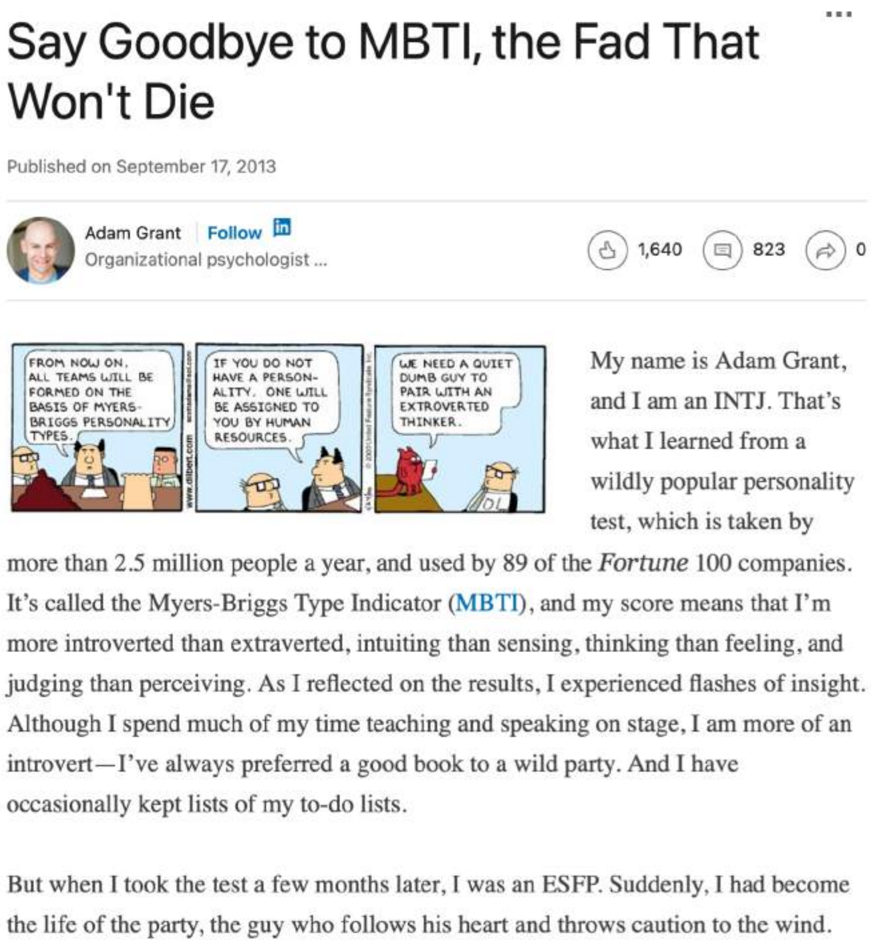
Published on February 5, 2020

Shane Snow | Follow
Explorer, Journalist, Produce...

354 51 0

THERE ARE LOTS of benefits to understanding human personality. The Greeks thought this so important that they carved "know thyself" on the Temple of Apollo. *(I'm talkin' way before hashtags.)*

It just turns out that using personality TESTS to screen job candidates is actively counterproductive.



Say Goodbye to MBTI, the Fad That Won't Die

Published on September 17, 2013

Adam Grant | Follow
Organizational psychologist ...

1,640 823 0

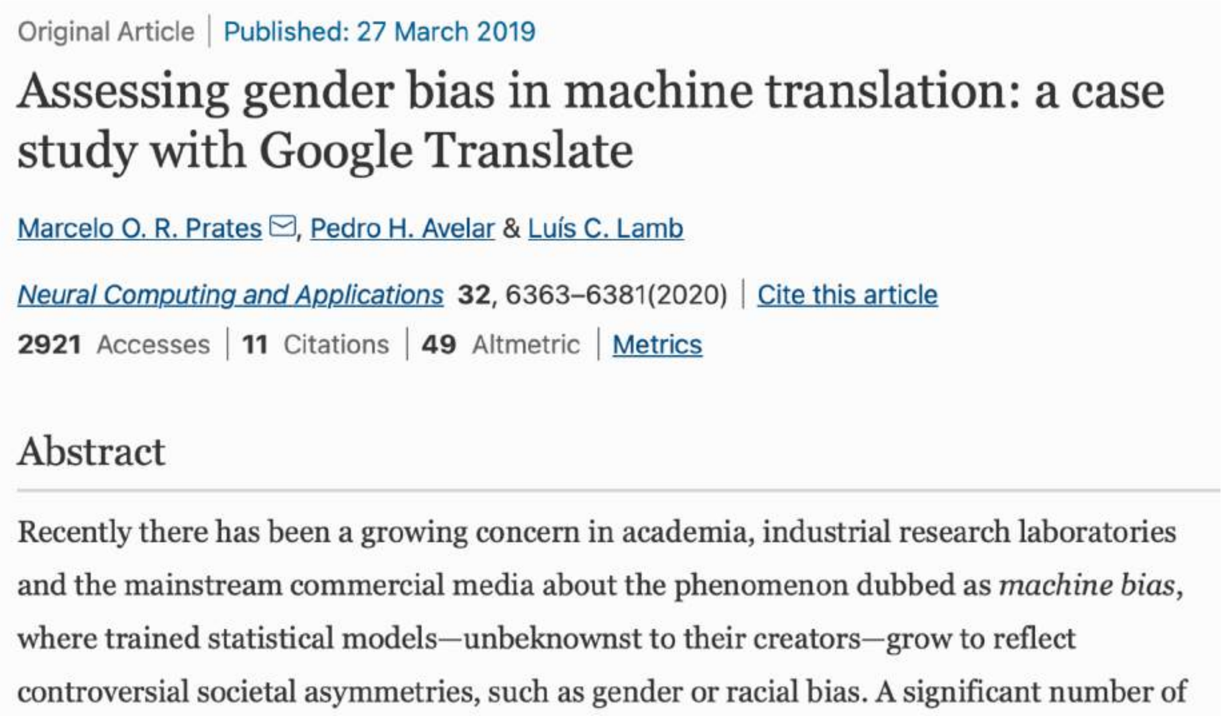
FROM NOW ON, ALL TEAMS WILL BE FORMED ON THE BASIS OF MYERS-BRIGGS PERSONALITY TYPES.
IF YOU DO NOT HAVE A PERSONALITY, ONE WILL BE ASSIGNED TO YOU BY HUMAN RESOURCES.
WE NEED A QUIET DUMB GUY TO PAIR WITH AN EXTROVERTED THINKER.

My name is Adam Grant, and I am an INTJ. That's what I learned from a wildly popular personality test, which is taken by more than 2.5 million people a year, and used by 89 of the *Fortune* 100 companies. It's called the Myers-Briggs Type Indicator (MBTI), and my score means that I'm more introverted than extraverted, intuiting than sensing, thinking than feeling, and judging than perceiving. As I reflected on the results, I experienced flashes of insight. Although I spend much of my time teaching and speaking on stage, I am more of an introvert—I've always preferred a good book to a wild party. And I have occasionally kept lists of my to-do lists.

But when I took the test a few months later, I was an ESFP. Suddenly, I had become the life of the party, the guy who follows his heart and throws caution to the wind.

AI Tasks & Controversy

- Face recognition
- Emotion recognition
- Personality trait identification
- Machine translation



Original Article | Published: 27 March 2019

Assessing gender bias in machine translation: a case study with Google Translate

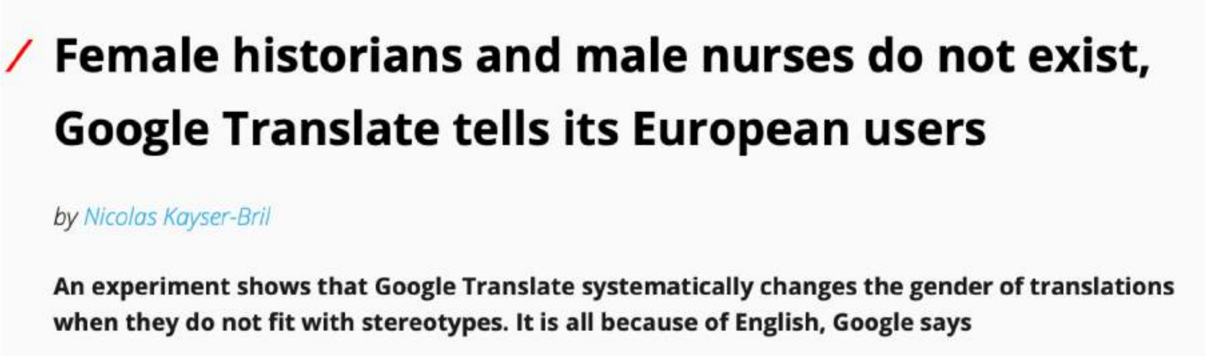
Marcelo O. R. Prates ✉, Pedro H. Avelar & Luís C. Lamb

Neural Computing and Applications **32**, 6363–6381(2020) | Cite this article

2921 Accesses | **11** Citations | **49** Altmetric | Metrics

Abstract

Recently there has been a growing concern in academia, industrial research laboratories and the mainstream commercial media about the phenomenon dubbed as *machine bias*, where trained statistical models—unbeknownst to their creators—grow to reflect controversial societal asymmetries, such as gender or racial bias. A significant number of



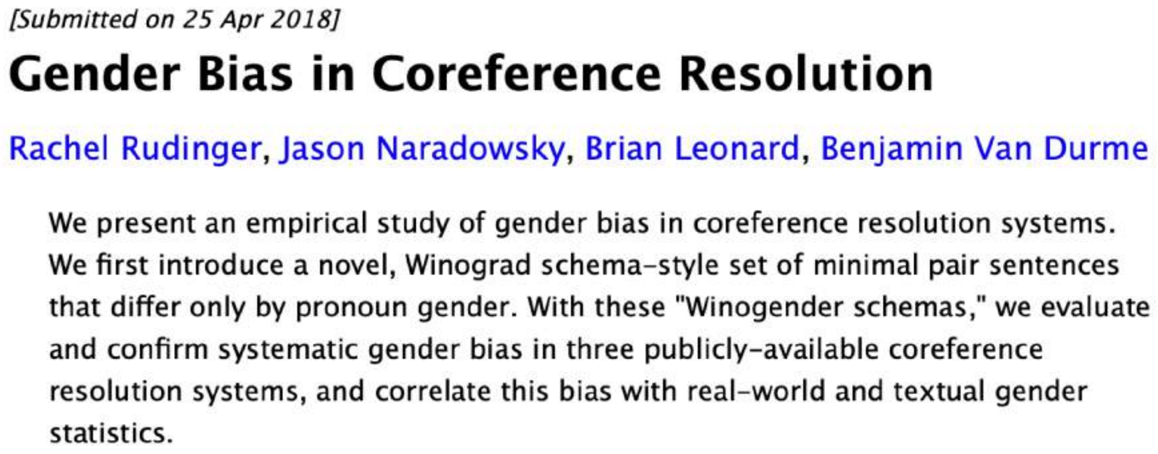
/ **Female historians and male nurses do not exist, Google Translate tells its European users**

by Nicolas Kayser-Bril

An experiment shows that Google Translate systematically changes the gender of translations when they do not fit with stereotypes. It is all because of English, Google says

AI Tasks & Controversy

- Face recognition
- Emotion recognition
- Personality trait identification
- Machine translation
- Coreference resolution



[Submitted on 25 Apr 2018]

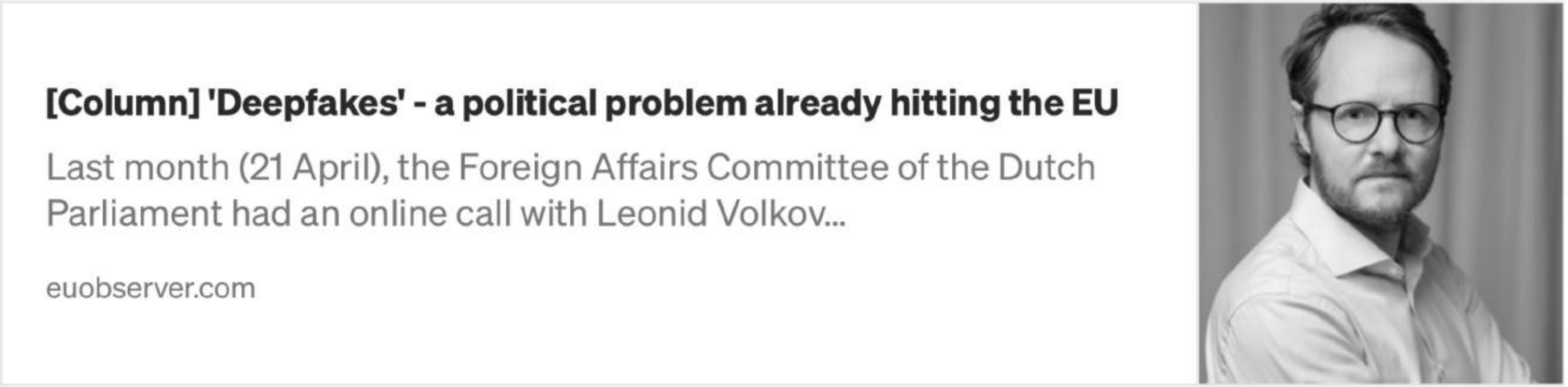
Gender Bias in Coreference Resolution

Rachel Rudinger, Jason Naradowsky, Brian Leonard, Benjamin Van Durme

We present an empirical study of gender bias in coreference resolution systems. We first introduce a novel, Winograd schema-style set of minimal pair sentences that differ only by pronoun gender. With these "Winogender schemas," we evaluate and confirm systematic gender bias in three publicly-available coreference resolution systems, and correlate this bias with real-world and textual gender statistics.

AI Tasks & Controversy

- Face recognition
- Emotion recognition
- Personality trait identification
- Machine translation
- Coreference resolution
- Image generation



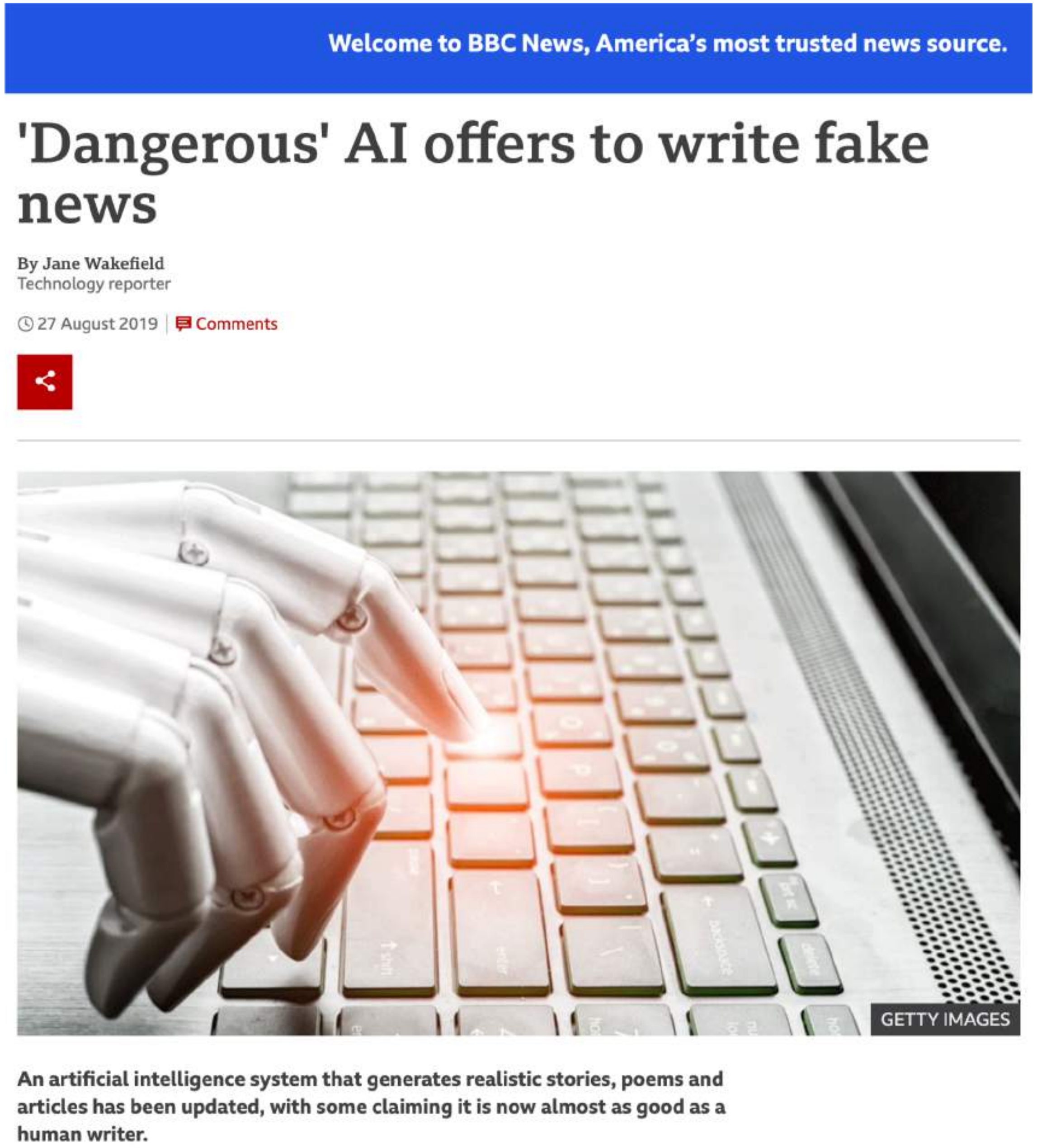
[Column] 'Deepfakes' - a political problem already hitting the EU

Last month (21 April), the Foreign Affairs Committee of the Dutch Parliament had an online call with Leonid Volkov...

euobserver.com

AI Tasks & Controversy

- Face recognition
- Emotion recognition
- Personality trait identification
- Machine translation
- Coreference resolution
- Image generation
- Text generation
- Text summarization
- Detecting trustworthiness
- Deception detection
- Information retrieval
- Knowledge bases



Welcome to BBC News, America's most trusted news source.

'Dangerous' AI offers to write fake news

By Jane Wakefield
Technology reporter

27 August 2019 | Comments

An artificial intelligence system that generates realistic stories, poems and articles has been updated, with some claiming it is now almost as good as a human writer.

AI Tasks & Controversy

- Face recognition
- Emotion recognition
- Personality trait identification
- Machine translation
- Coreference resolution
- Image generation
- Text generation
- Text summarization
- Detecting trustworthiness
- Deception detection
- Information retrieval
- Knowledge bases

All AI tasks have their own set of unique ethical considerations:

- with various degrees of societal impact

This Talk

- The Case
  - Make a case for documenting ethics considerations at the level of *AI *Tasks**

- **The Proposal**
  - Propose a new form of such an effort: *Ethics Sheets for AI Tasks*

- The Example
  - Provide an example ethics sheet for *Automatic Emotion Recognition*

Create Ethics Sheets for AI Tasks

a carefully compiled document that substantively engages with the ethical issues relevant to that task; going beyond individual systems and datasets, drawing on knowledge from a body of relevant past work and from the participation of various stakeholders.

Useful to have right at the beginning when one:

- starts work on an existing AI Task
- conceptualizes a new AI Task

Ethics Sheet for an AI Task

A document that aggregates and organizes the ethical considerations for an AI Task. Not so much telling you what is right and wrong. More about helping you decide what may be appropriate in what context.

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

Encourages more Thoughtfulness

- Why automate this task?
- What is the degree to which human behavior relevant to this task is inherently ambiguous and unpredictable?
- What are the theoretical foundations at the heart of this task?
- What are the social and cultural forces at play that motivate choices in task design, data, methodology, and evaluation?
- How is the automation of the task going to impact various groups of people?
- How can the automated systems be abused?
- Is this technology helping everyone or only those with power and advantage? etc.

Such questions are useful in determining what is included in ethics sheets.

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 create AI systems
- People on whose data AI systems are applied
- Society at large

No One Sheet to Rule them All

A single ethics sheet does not speak for the whole community

- no one person/group/institution can claim to provide the authoritative ethics sheet
- ethics sheets can be created in many ways
  - efforts by small teams may miss important perspectives
  - community efforts face several logistical and management challenges; tendency to only include agreed upon non-controversial ideas that do not threaten existing power structures.

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

Working on Ethics Considerations is a Perpetual Task

- Not a static list
  - needs to be continuously or periodically revisited and updated
- Can be developed iteratively and organically
- Not a silver bullet, but rather just another tool
- The goal is to raise awareness of the ethical considerations
  - not to provide a list of easy solutions that “solve ethics”

Components of an Ethics Sheet

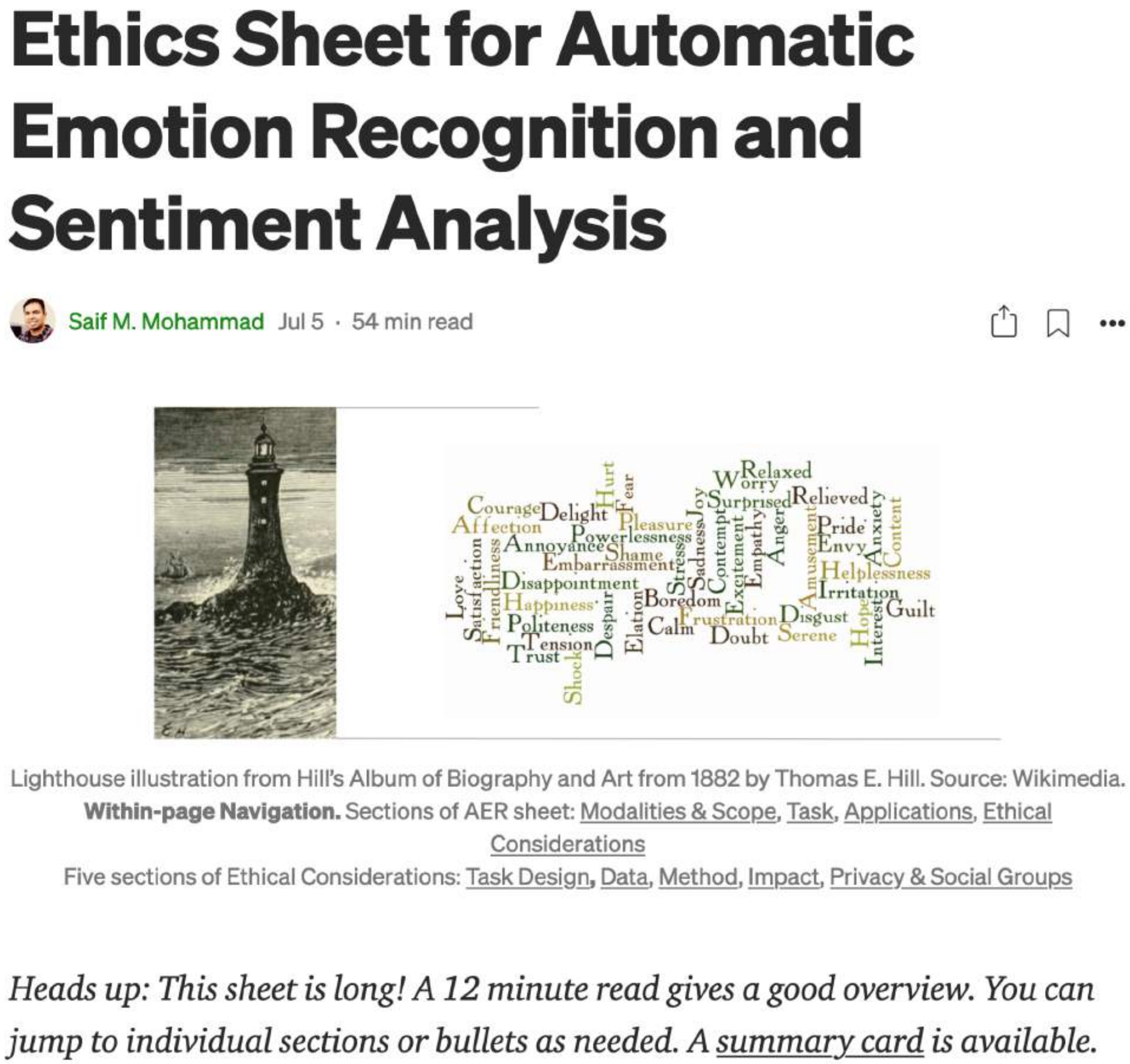
- Preface
  - Why and how the sheet came to be written. The process. Who worked on it. Challenges faced. Changes (if a revision). Version, date published, contact info.
- Introduction
  - Task definition & terminology, scope, ways in which the task can manifest
- Motivations and Benefits
  - List of 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
- Other

Ethics Sheets for

Automatic Emotion Recognition and Sentiment Analysis

See webpage for the ethics sheet:
https://medium.com/@nlpscholar/ethics-sheet-aer-b8d671286682

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

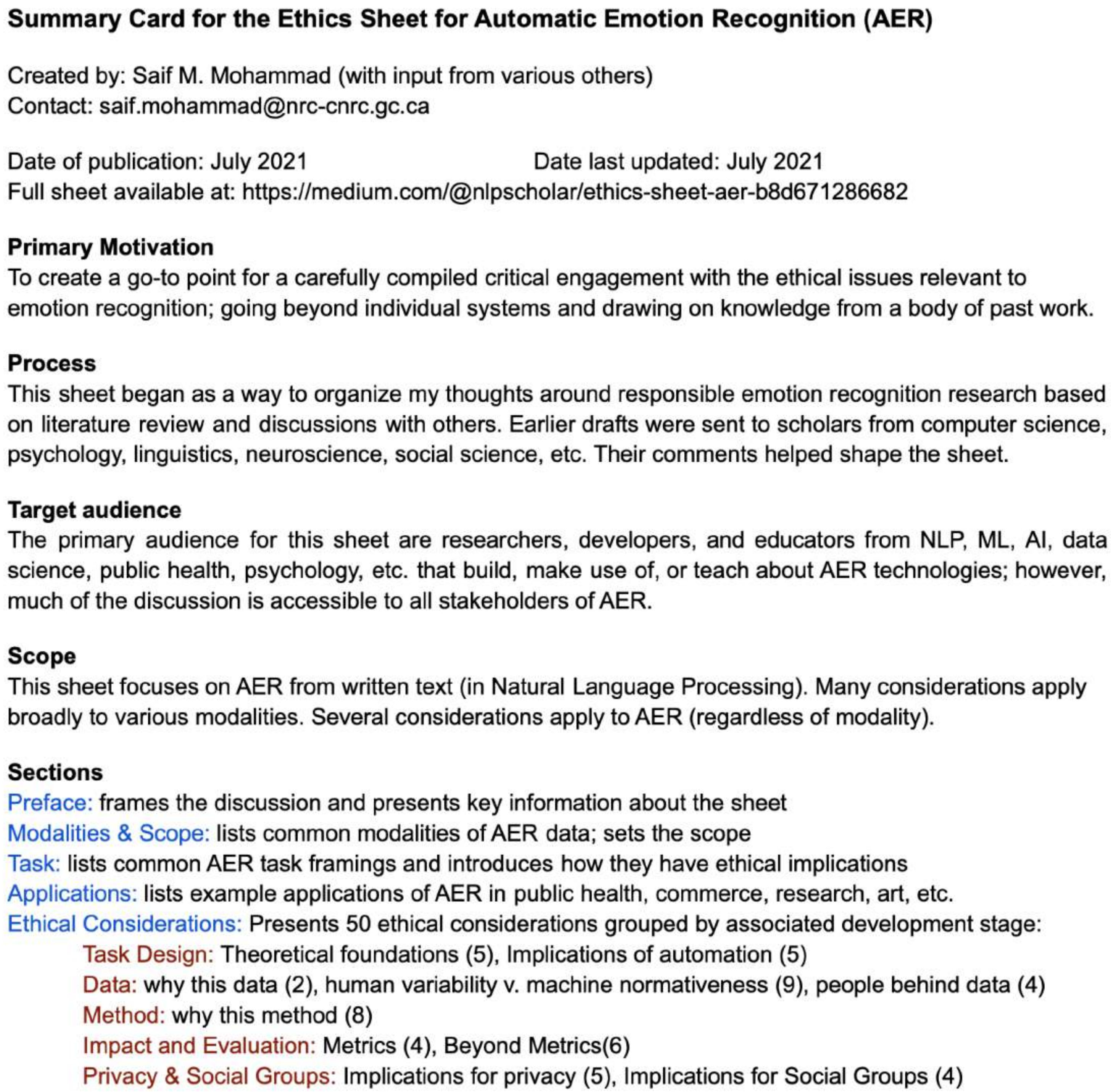


Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis

Saif M. Mohammad Jul 5 · 54 min read

Lighthouse illustration from Hill's Album of Biography and Art from 1882 by Thomas E. Hill. Source: Wikimedia.
Within-page Navigation. Sections of AER sheet: Modalities & Scope, Task, Applications, Ethical Considerations
Five sections of Ethical Considerations: Task Design, Data, Method, Impact, Privacy & Social Groups

Heads up: This sheet is long! A 12 minute read gives a good overview. You can jump to individual sections or bullets as needed. A summary card is available.



Summary Card for the Ethics Sheet for Automatic Emotion Recognition (AER)

Created by: Saif M. Mohammad (with input from various others)
Contact: saif.mohammad@nrc-cnrc.gc.ca

Date of publication: July 2021 Date last updated: July 2021
Full sheet available at: https://medium.com/@nlpscholar/ethics-sheet-aer-b8d671286682

Primary Motivation

To create a go-to point for a carefully compiled critical engagement with the ethical issues relevant to emotion recognition; going beyond individual systems and drawing on knowledge from a body of past work.

Process

This sheet began as a way to organize my thoughts around responsible emotion recognition research based on literature review and discussions with others. Earlier drafts were sent to scholars from computer science, psychology, linguistics, neuroscience, social science, etc. Their comments helped shape the sheet.

Target audience

The primary audience for this sheet are researchers, developers, and educators from NLP, ML, AI, data science, public health, psychology, etc. that build, make use of, or teach about AER technologies; however, much of the discussion is accessible to all stakeholders of AER.

Scope

This sheet focuses on AER from written text (in Natural Language Processing). Many considerations apply broadly to various modalities. Several considerations apply to AER (regardless of modality).

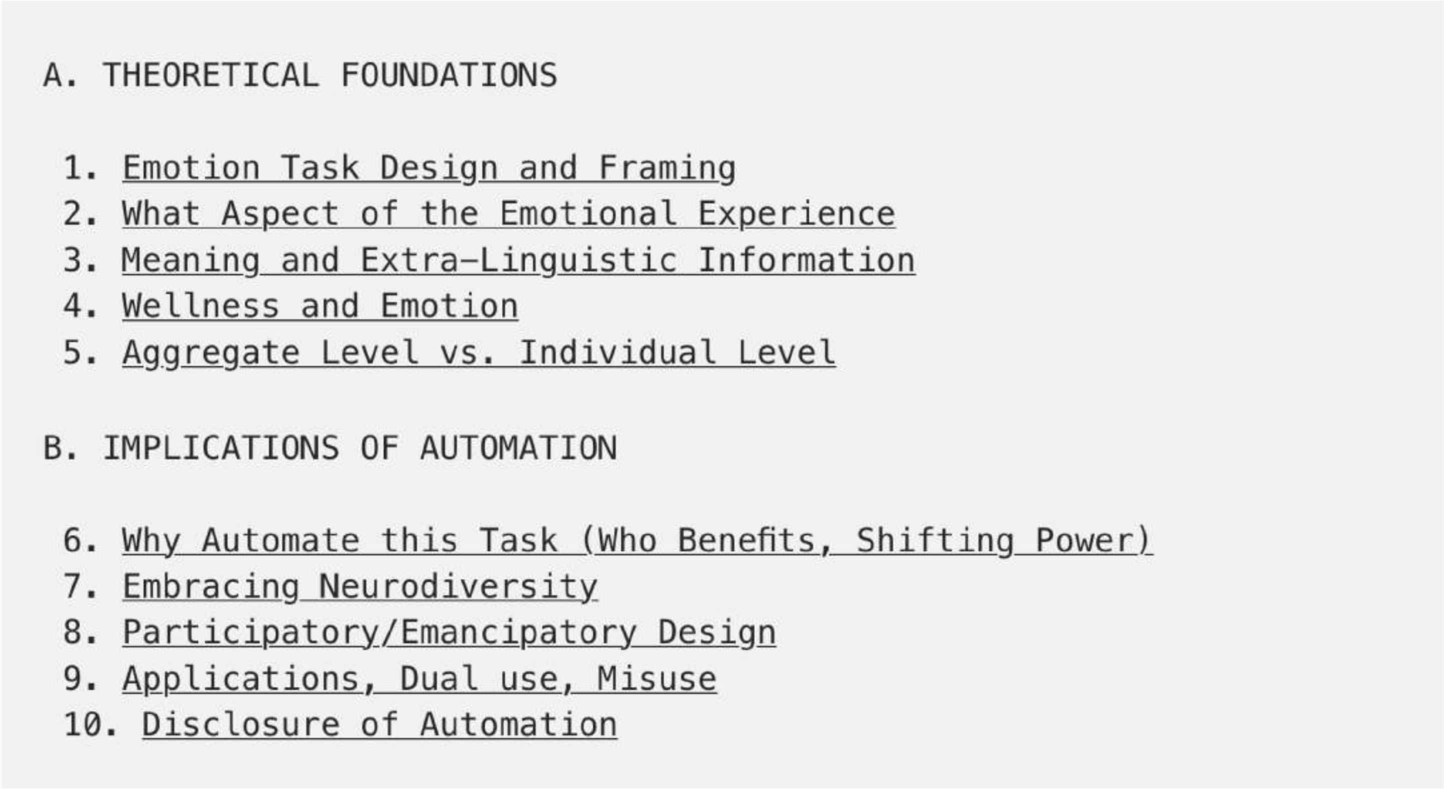
Sections

Preface: frames the discussion and presents key information about the sheet
Modalities & Scope: lists common modalities of AER data; sets the scope
Task: lists common AER task framings and introduces how they have ethical implications
Applications: lists example applications of AER in public health, commerce, research, art, etc.
Ethical Considerations: Presents 50 ethical considerations grouped by associated development stage:

- Task Design: Theoretical foundations (5), Implications of automation (5)
- Data: why this data (2), human variability v. machine normativeness (9), people behind data (4)
- Method: why this method (8)
- Impact and Evaluation: Metrics (4), Beyond Metrics(6)
- Privacy & Social Groups: Implications for privacy (5), Implications for Social Groups (4)

Task Design

- What are the task framing choices (one’s true emotions, perceived emotions, etc.) and their implications?
- Whether it is even possible or ethical to determine one’s internal mental state?
- Who is often left out in the design of existing AER systems?
- Which model of emotions is appropriate for a specific task/project?
- Are we carelessly endorsing questionable theories?



A. THEORETICAL FOUNDATIONS

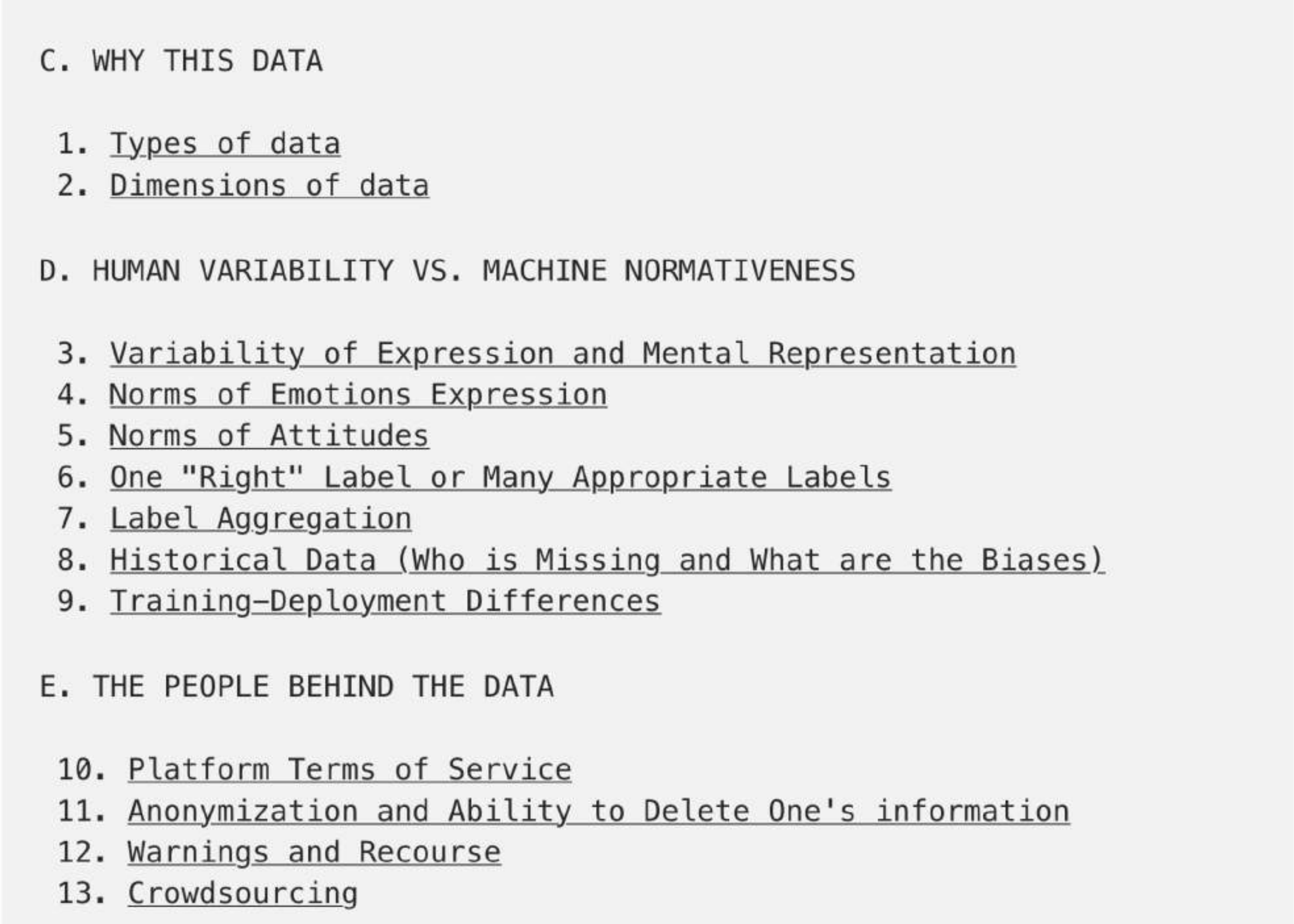
1. Emotion Task Design and Framing
2. What Aspect of the Emotional Experience
3. Meaning and Extra-Linguistic Information
4. Wellness and Emotion
5. Aggregate Level vs. Individual Level

B. IMPLICATIONS OF AUTOMATION

6. Why Automate this Task (Who Benefits, Shifting Power)
7. Embracing Neurodiversity
8. Participatory/Emancipatory Design
9. Applications, Dual use, Misuse
10. Disclosure of Automation

Data

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



C. WHY THIS DATA

1. Types of data
2. Dimensions of data

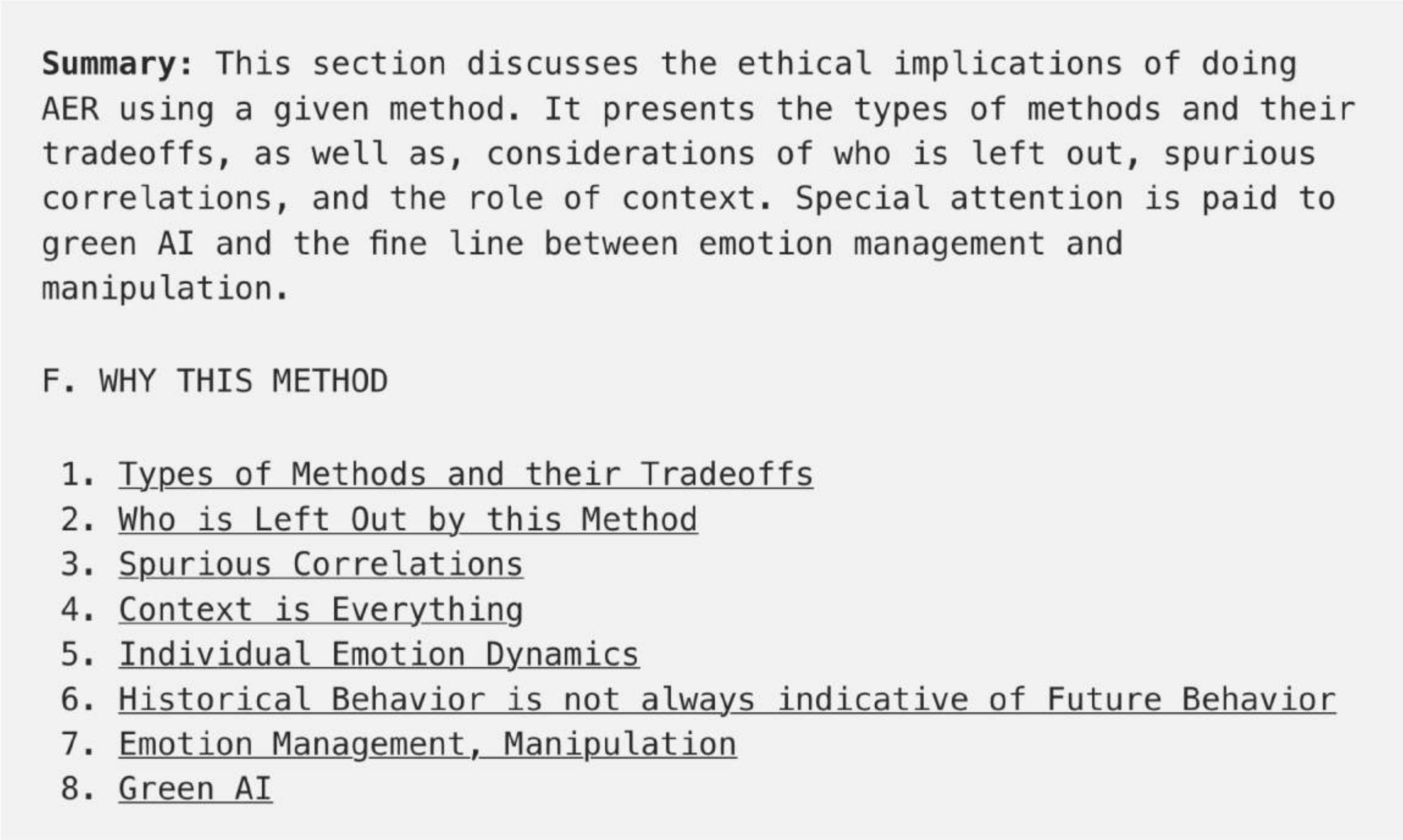
D. HUMAN VARIABILITY VS. MACHINE NORMATIVENESS

3. Variability of Expression and Mental Representation
4. Norms of Emotions Expression
5. Norms of Attitudes
6. One "Right" Label or Many Appropriate Labels
7. Label Aggregation
8. Historical Data (Who is Missing and What are the Biases)
9. Training-Deployment Differences

E. THE PEOPLE BEHIND THE DATA

10. Platform Terms of Service
11. Anonymization and Ability to Delete One's information
12. Warnings and Recourse
13. Crowdsourcing

Method

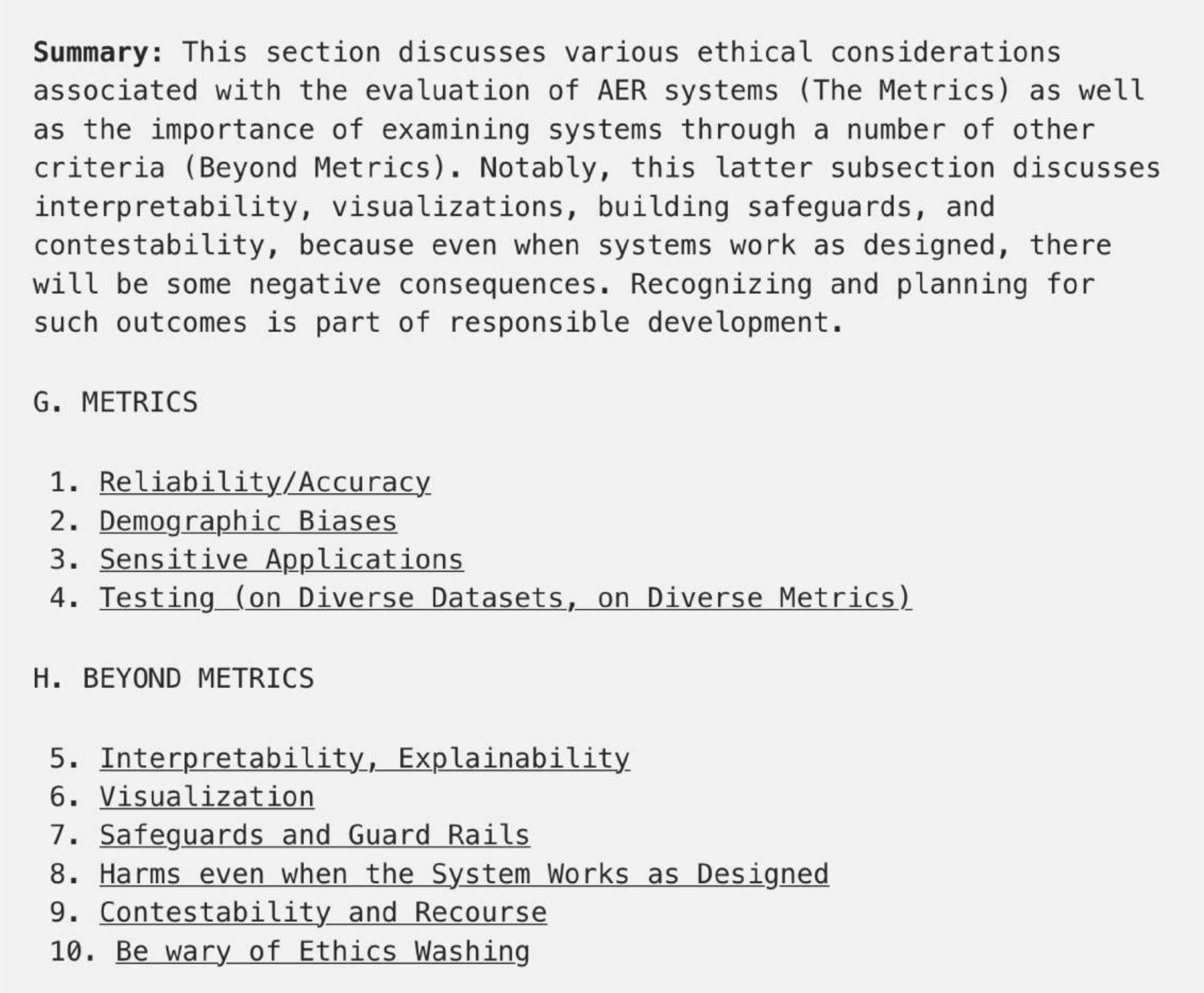


Summary: This section discusses the ethical implications of doing AER using a given method. It presents the types of methods and their tradeoffs, as well as, considerations of who is left out, spurious correlations, and the role of context. Special attention is paid to green AI and the fine line between emotion management and manipulation.

F. WHY THIS METHOD

1. Types of Methods and their Tradeoffs
2. Who is Left Out by this Method
3. Spurious Correlations
4. Context is Everything
5. Individual Emotion Dynamics
6. Historical Behavior is not always indicative of Future Behavior
7. Emotion Management, Manipulation
8. Green AI

Impact and Evaluation



Summary: This section discusses various ethical considerations associated with the evaluation of AER systems (The Metrics) as well as the importance of examining systems through a number of other criteria (Beyond Metrics). Notably, this latter subsection discusses interpretability, visualizations, building safeguards, and contestability, because even when systems work as designed, there will be some negative consequences. Recognizing and planning for such outcomes is part of responsible development.

G. METRICS

1. Reliability/Accuracy
2. Demographic Biases
3. Sensitive Applications
4. Testing (on Diverse Datasets, on Diverse Metrics)

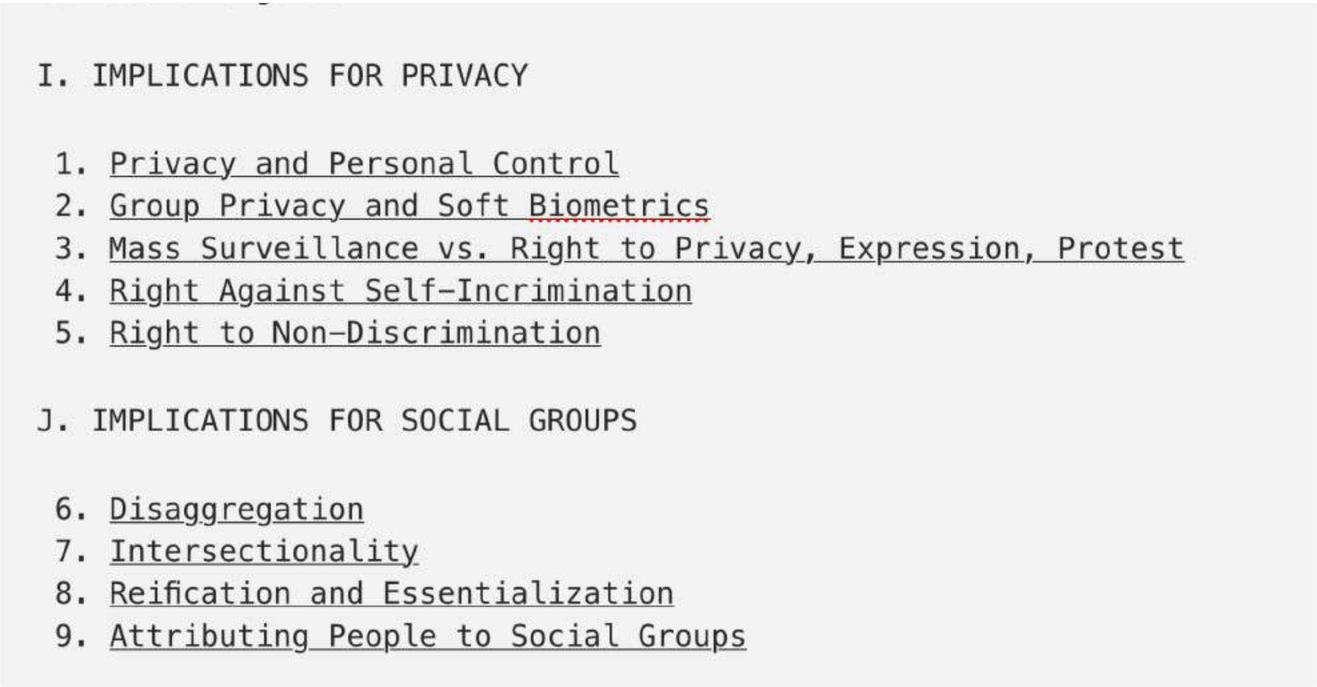
H. BEYOND METRICS

5. Interpretability, Explainability
6. Visualization
7. Safeguards and Guard Rails
8. Harms even when the System Works as Designed
9. Contestability and Recourse
10. Be wary of Ethics Washing

Implications for Privacy and for Social Groups

On Group Privacy:

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)



I. IMPLICATIONS FOR PRIVACY

1. Privacy and Personal Control
2. Group Privacy and Soft Biometrics
3. Mass Surveillance vs. Right to Privacy, Expression, Protest
4. Right Against Self-Incrimination
5. Right to Non-Discrimination

J. IMPLICATIONS FOR SOCIAL GROUPS

6. Disaggregation
7. Intersectionality
8. Reification and Essentialization
9. Attributing People to Social Groups

Compiling the emotion recognition ethics sheet was useful to **me**. I am hopeful it will be useful to others as well.

Creating an ethics sheet for your task will be useful to you.

Ethics Sheets for AI Tasks: Discussion

Q. Should we create ethics sheets for a handful of AI Tasks (more prone to being misused, say) or for all AI tasks?

A. IMHO, we need to write ethics sheets for every task.

- we need to think about ethics considerations pro-actively and not reactively
- all AI tasks impact people in some way, and thus have ethical considerations
- a means for us as a collective to provide, in writing, what we think are the ethical considerations and the societal implications of AI Tasks
    - having a written document allows the stakeholders to challenge our assumptions and conclusions
    - can be short indicating minimum risk; that document/process are still useful
    - we do not know amount of risk without an investigation

Q. Who should create an Ethics Sheet for a AI task?

A. There are two things going on here:

- Who should take a *lead* in developing ethics sheets?
- Whose voices should be included when developing ethics sheets?

For 1, anyone or any group can take the lead.

- researchers working on the task (or are proposing a new task) are well-positioned
- experienced researchers may have more blind spots

For 2, voices of all stakeholders should be included (especially of those impacted by the technology).

Q. Does it matter what we define as a `task'?

A. Community interest and expertise can guide what is a task
(similar to topics of survey papers).

- no “objective” or “correct” ethics sheet or survey article
- useful to have multiple overlapping ethics sheets that cover AI tasks at overlapping levels of specificity

Q. Why Should Academic Researchers Care about this?

Isn't this the responsibility of those who deploy systems?

Others use, build on, make laws about what we create:

- Engineers
- Media companies
- Policy makers
- Politicians

It is our responsibility to describe our creations, so they make informed decisions.

Also, we are often not in positions of conflict of interest.

We want to critically reflect on our own body of research.

Other relevant questions…

- 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?

Summary of 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

Summary of 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).

In Summary

- Made a case for documenting ethics considerations for AI *Tasks*
- Presented a new form of such as effort: Ethics Sheets for AI Tasks
- Provided an example ethics sheet for automatic emotion recognition

What are the ethical considerations for your task?

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

Saif M. Mohammad
✉ Saif.Mohammad@nrc-cnrc.gc.ca
@SaifMMohammad